The clinical potential of artificial intelligence in early detection of lung cancer

Høstgaard Bang Henriksen, Margrethe

DOI:
10.21996/feab38bd-ecdf-4b1d-9cc6-5e8e91a483bd

Publication date:
2025

Document version:
Final published version

The clinical potential of artificial intelligence in early detection of lung cancer

Margrethe Høstgaard Bang Henriksen, MD

Department of Oncology, Lillebaelt Hospital,

University Hospital of Southern Denmark, Vejle

Department of Regional Health Research

Faculty of Health Sciences

University of Southern Denmark

2024

Supervisors

Torben Frøstrup Hansen (Main supervisor)

Professor, MD, PhD, DMSc

Department of Oncology, Lillebaelt Hospital, University Hospital of Southern Denmark, Vejle,

Denmark

Ole Hilberg (Assistant supervisor)

Professor, MD, DMSc,

Department of Medicine, Lillebaelt Hospital, University Hospital of Southern Denmark, Vejle, Denmark

Claus Lohman Brasen (Assistant supervisor)

MD, PhD

Department of Biochemistry and Immunology, Lillebaelt Hospital, University Hospital of Southern Denmark, Kolding, Denmark

Lars Henrik Jensen (Assistant supervisor)

Associate Professor, MD, Ph.D,

Department of Oncology, Lillebaelt Hospital, University Hospital of Southern Denmark, Vejle, Denmark

Assessment committee

Pernille Just Vinholt (Chairman)

Professor, MD, PhD

Department of Clinical Biochemistry

Odense University Hospital

Odense, Denmark

Oluf Dimitri Røe

Professor, MD, PhD

Department of Clinical Research and Molecular Medicine, Trondheim, Norway

Cancer Clinic, Levanger Hospital, Levanger, Norway

Clinical Cancer Research Center, Department of Clinical Medicine, Aalborg University Hospital

Aalborg, Denmark

Janus Laust Thomsen

Clinical Professor

Department of Clinical Medicine

Aalborg University Hospital

Aalborg, Denmark

Preface

As I near the completion of my PhD, I find myself on the entire six-year journey that began in 2018 when I first started planning this project. It has been a long and eventful path, starting from a point where I knew almost nothing about the intersection of AI and cancer. However, I embraced the challenge, and as I progressed up the learning curve, my curiosity deepened, ultimately leading me to focus on lung cancer detection.

The original study plan quickly became outdated as I gained a better understanding of the clinical perspectives and the availability of data. Although the main focus remained on lung cancer detection, the specific sub-projects and related publications evolved based on new knowledge, inspiration from collaborators, and my growing interest in the field.

The early phase of my PhD was marked by fund raising, PhD courses, obtaining study approvals, applying for access to different data sources, and becoming well-acquainted with the competent staff at the regional data warehouse and the RKKP databases. Additionally, I was fortunate to establish a strong partnership with the Mærsk Mc-Kinney Møller Institute at the University of Southern Denmark (SDU). Here, I collaborated with several master's students and their supervisors. Weekly meetings and courses in data science gave me valuable insights into the workings of a data scientist, and I hope I was able to contribute with meaningful clinical insights from my perspective.

During the second phase of my PhD, the collaboration with SDU expanded to include work with large language models to predict smoking status - an interesting project that we continue to explore. I also had the privilege of being introduced to the skilled colleagues at Maastricht University, where I spent time with their large data science team at the Maastricht Radiation Oncology clinic. This experience led to another collaboration exploring other AI methods with our data. Around this time, Roche expressed interest in our work, which led to a partnership focused on Danish validation of an international lung cancer prediction model.

In the final phase of my PhD, my efforts were primarily directed toward completing the various projects and publications I had initiated as well as planning my thesis and postdoc. I also became involved with an EU project at the Department of Oncology, and I look forward to continuing my work in AI, big data, and prediction models with our local research team.

Throughout this journey, I have been fortunate to receive interest and support from the Department of Oncology, the hospital management, colleagues, collaborators, friends, family and even the media. I have had the opportunity to present my research on several occasions, both to colleagues and the general public. I take great pleasure in disseminating knowledge, particularly in making data science accessible to clinicians. I hope that this thesis reflects that passion, and I hope you find as much enjoyment in reading it as I did in writing it.

Acknowledgements

During my PhD journey I have received invaluable support at every stage, from the initial financial backing to data collection, analysis, and supervision.

I am particularly grateful to the funds and institutions that made this project possible: the Region of Southern Denmark, the University of Southern Denmark, the Danish Research Center for Lung Cancer, the Dagmar Marshall Foundation, the Beckett Foundation, the Lilly and Herbert Hansen Foundation, and the Hede Nielsen Family Foundation.

During data collection, I received invaluable support from the regional data warehouse, OPEN, and RKKP, with special thanks to Rune Pedersen, Sune Welling, Kristian Kvist, Harald Hammersøi, and Henriette Engberg. I also want to acknowledge Martin Ask Klausholt and Elise Humerfelt for their help in annotating large volumes of free text, and I appreciate the legal guidance provided by Randi Bilberg, Claus Kvist Hansen, and SDU RIO.

For data analysis, I am thankful to Ricco Noel Hansen Flyckt, Louise Sjodsholm, Jeppe Drue Knudsen, Ali Ebrahimi, Abdolrahman Peimankar and Uffe Kock Wiil from the Mærsk McKinney Møller Institute as well as the collaborative support from Florian Van Daalen, Leonard Wee, and Inigo Bermejo from the MAASTRO Clinic. It has been truly inspirational to work with you all, and I look forward to future collaborations.

My sincere thanks go to my supervisors, Torben Frøstrup Hansen, Lars Henrik Jensen, Claus Brasen, and Ole Hilberg for their guidance across oncology, biochemistry, and internal medicine. You have given me the freedom to manage my own project while always being available for guidance and support when needed. I am also thankful for the advice I received from Anders Løkke and Morten Borg in pulmonology as well as from Jens Søndergaard in general practice. I also appreciate the support from the lung cancer fast-track clinic in Vejle. Special thanks to Karin Larsen for language supervision and managing formalities, and to the Department of Oncology and Lars Henrik Jensen for their belief in this project.

Lastly, I want to thank my PhD colleagues for fostering a supportive environment and my family for their unwavering support and encouragement all the way through this journey.

List of included articles

Article I

A collection of multiregistry data on patients at high risk of lung cancer - a Danish retrospective cohort study of nearly 40,000 patients, Henriksen MB, Hansen TF, Jensen LH, Brasen CL, Peimankar A, Ebrahimi A, Wiil UK, Hilberg O. Translational Lung Cancer Research. 2023 Dec;12(12):2392-2411. Doi:10.21037/tlcr-23-495[1]

Article II

Pulmonologist-Level lung cancer detection based on standard blood test results and smoking status using an explainable machine learning approach, Flyckt RNH[‡], Sjodsholm L[‡], Henriksen MB[‡], Brasen CL, Ebrahimi A, Hilberg O, Hansen TF, Wiil UK, Jensen LH, Peimankar A. Scientific Reports, accepted for publication. 2024[2]

Article III

Lung cancer detection using Bayesian networks: A retrospective development and validation study on a Danish population of high-risk individuals, Henriksen MB[‡], Van Daalen F[‡], Wee L, Hansen TF, Jensen LH, Brasen CL, Hilberg O, Bermejo I. Cancer Medicine, accepted for publication. 2024

Article IV

A Bayesian Network approach to lung cancer screening: Assessing the impact of data quantity, quality and the combination of variables from Danish electronic health records, Van Daalen F[‡], Henriksen MB[‡], Hansen TF, Jensen LH, Brasen CL, Hilberg O, Andersen MK, Humerfelt E, Wee L, Bermejo I. Cancers, under peer review. 2024

Article V

Lung Cancer among outpatients with Chronic Obstructive Pulmonary Disease - A seven-year cohort Study, Henriksen MB, Hansen TF, Jensen LH, Brasen CL, Borg M, Hilberg O, Løkke A.ERJ Open Research. 2024; 10: 64–2024. Doi: 10.1183/23120541.00064-2024[3]

[‡] Shared first authorship

Related articles not included in this thesis

Identification of patients' smoking status using an explainable AI approach: a Danish electronic health records case study, Ebrahimi A.[ǂ], Henriksen M.B.[ǂ], Brasen C.L. *et al.* BMC Med Res Methodology. 2024, 114 (2024). Doi: 10.1186/s12874-024-02231-4 [4]

[ǂ] Shared first authorship

Table of contents

01 Abbreviations

AI	Artificial intelligence
ALAT	Alanine transaminase
ATC	Anatomic therapeutic chemical code
AUC	Area under the receiver operating characteristic curve
BN	Bayesian network
CCI	Charlson comorbidity index
CRP	C-reactive protein
DAG	Directed acyclic graph
DES	Dynamic ensemble selection
DLCST	Danish lung cancer screening trial
DrCOPD	Danish registry of chronic obstructive pulmonary disease
EHR	Electronic health records
FEV_1	Forced expiratory volume in 1 second
FPR	False positive rate
FNR	False negative rate
GDPR	General data protection regulation
HUNT	Helseundersøkelsen i Nord-Trøndelag
ICS	Inhaled corticosteroids
INR	International normalized ratio
IQR	Interquartile range
LABA	Long-acting β-agonists
LCRAT	Lung cancer risk assessment tool
LAMA	Long-acting muscarinic antagonists

LDCT	Low-dose computed tomography
ICD-10	International classification of diseases, version 10
LC	Lung cancer
LDH	Lactate dehydrogenase
LLP	Liverpool Lung Project
LR	Logistic regression
ML	Machine learning
MCAR	Missing completely at random
MRC	Medical Research Council
NCCN	National comprehensive cancer network
NLST	National Lung Screening Trial
NELSON	Nederlands-Leuvens Longkanker Screenings Onderzoek
PCA	Principal component analysis
PET	Positron Emission Tomography
PPV	Positive predictive value
ROC	Receiver operating characteristic
SD	Standard deviation
SHAP	Shapley Additive exPlanations
SVM	Support vector machine
TPR	True positive rate
TNR	True negative rate
USPSTF	United States Preventive Services Task Force
4-IN-THE-LUNG-RUN	Towards INdividually tailored INvitations, screening INtervals, and INtegrated co-morbidity reducing strategies in lung cancer screening

02 English summary

Lung cancer (LC) is currently the leading cause of cancer-related deaths, highlighting the critical necessity for early detection, which is essential for providing curative treatment. While screening for LC is gradually introduced through pilot studies across various countries, discussions persist regarding the optimal selection criteria. Numerous studies have highlighted the superiority of individual prediction models over the widespread categorical standard criteria based solely on age and smoking intensity.

The overall aim of this thesis was to explore and refine LC detection models based on artificial intelligence (AI) utilizing data obtained from clinical health records and registries. The issue was addressed from several angles, resulting in the incorporation of five articles in this thesis.

The first four studies revolved around data derived from a high-risk cohort of patients evaluated in the LC fast-track clinics in the Region of Southern Denmark. Extensive clinical and laboratory data were collected from this cohort of nearly 40,000 individuals, including 25% of which were LC patients. Associations between data variables and LC status were examined in **Article I**, laying the groundwork for subsequent prediction models. The initial findings let to the usage of smoking and laboratory data in the development of prediction models employing both a machine learning (ML) approach (**Article II**) and a Bayesian Networks (BN) approach (**Article III**). The ML model exhibited similar performance to the BN approach with a mean area under the receiver operating characteristic (ROC) curve (AUC) of 0.77 compared to AUC 0.76, and both with a sensitivity of 21% at a fixed specificity of 95%. The ML model identified smoking status, lactate dehydrogenase, age and plasma calcium levels as the most important factors for detection of LC. The BN model demonstrated performance robustness when introduced to missing data (up to 30%), a notable advantage when working with clinical data analysis.

Additional data types such as symptoms at diagnosis, comorbidities, and medication were integrated into an expanded BN model, investigating whether a more comprehensive dataset could enhance model performance (**Article IV**). The best-performing model achieved an AUC of 0.79 and was developed using comorbidity, laboratory results, and smoking data on a relatively large dataset with 21% missing variables. Additionally, a model developed on a small but complete dataset proved to be stable when applied to larger datasets with up to 39% missing data, indicating its applicability in individuals with incomplete data. While laboratory results and smoking status were the strongest predictors of LC, comorbidity (including data on medication and data from general practice) and symptoms at diagnosis appeared to be the least informative.

While the first four studies focused solely on high-risk patients, we aimed to extrapolate these findings to a potentially lower-risk population eligible for LC screening. Therefore, we assessed the risk of LC and the overlap with LC fast-track clinics among chronic obstructive pulmonary disease (COPD) outpatients (**Article V**). Within this cohort, we observed a 5% risk of LC, surpassing the risk in the general population more than tenfold. Importantly, LC patients with COPD were diagnosed at an earlier stage than LC patients without COPD. Additionally, 18% of COPD outpatients were referred to LC diagnostics at some point. While this high referral rate may be due to increased medical attention, it suggests potential benefits of a regular and systematic screening approach for these patients.

The insights and methodology outlined in this thesis serve as foundational elements for our ongoing research, which aims to integrate risk models into a practical clinical screening context. A highly effective model capable of early-stage LC prediction will enhance screening efficacy and promote early detection, ultimately leading to improved survival rates.

03 Dansk lægmands-resumé

Lungekræft (LC) er i øjeblikket den kræftsygdom der fører til flest dødsfald, og det er afgørende, at sygdommen opdages tidligt, så man kan tilbyde helbredende behandling. Screening for LC er derfor gradvist ved at blive afprøvet og indført i form af pilotprojekter i en lang række lande. Der hersker dog stadig uklarhed om udvælgelseskriterierne. Flere studier viser, at individuelle risikomodeller er bedre end de mere simple standardkriterier, der kun tager højde for alder og rygning.

Formålet med denne afhandling var at undersøge og udvikle modeller til at opdage LC ved brug af kunstig intelligens (AI) baseret på data fra patientjournaler, laboratorieresultater og registre. Emnet blev belyst fra flere vinkler med fem artikler inkluderet i afhandlingen.

De første fire studier inkluderer data fra ca. 40.000 individer udredt på mistanke om LC i pakkeforløb i Region Syddanmark. De udgør en højrisikogruppe, hvoraf 25% vist sig at have LC. I **Artikel I** blev der set på sammenhængen mellem de forskellige kliniske og biokemiske data og om patienterne havde LC, hvilket dannede grundlaget for senere prædiktionsmodeller. På baggrund af de første resultater blev rygning og laboratoriedata brugt til at udvikle prædiktionsmodeller baseret på både maskinlæring (ML) (**Artikel II**) og en anden metode kaldet Bayesianske Netværk (BN) (**Artikel III**). Modellernes prædiktionsevne var sammenlignelig med en sensitivitet på 21% for begge modeller ved en specificitet på 95%. ML-modellen var i stand til at identificerede rygning, et leverenzym, alder og calciumniveau som de vigtigste parametre. BN-modellen var i stand til at prædiktere stabilt, selv når der manglede op til 30% af dataene, hvilket er en stor fordel når man arbejder med kliniske data, der sjældent er komplette.

I en udvidet BN-model inddragede vi også andre datakilder, heriblandt symptomer ved diagnose og sygdoms- og medicinhistorik (**Artikel IV**). En af de bedste modeller blev udviklet på gruppen af individer med alle datatyper tilgængelige, og denne model kunne efterfølgende bruges på en gruppe, hvor mængden af tilgængelig data var mere begrænset. De vigtigste datatyper for modellerne var rygning og laboratorieresultater, mens oplysninger om sygdomshistorik og symptomer havde mindre betydning.

Baseret på erfaringerne fra studie I-IV omhandlende højrisikopatienter ønskede vi at fokusere på en gruppe med lavere forekomst af LC, som potentielt kunne være egnet til LC-screening. Derfor undersøgte vi risikoen for LC blandt patienter fulgt i hospitalsregi med kronisk obstruktiv lungesygdom (KOL) og ligeledes hyppigheden af individer udredt i LC pakkeforløb blandt denne gruppe (**Artikel V**). KOL-patienterne viste sig at have en risiko for LC på 5%, hvilket er mere end ti gange højere end den generelle befolkning. Desuden blev LC-patienter med KOL diagnosticeret i et tidligere stadie end LC-patienter uden KOL.

Cirka hver femte KOL-patient blev også udredt for lungekræft i pakkeforløb, hvilket kan skyldes øget lægelig opmærksomhed, men det indikerer, at regelmæssig screening for LC kan komme disse patienter til gavn.

Resultaterne og metoderne i denne afhandling danner grundlag for vores videre forskning, som har til formål at integrere risikomodeller i klinisk screeningssammenhæng. En velfungerende og præcis model, der kan prædiktere LC i et tidligt stadie, vil potentielt kunne fremme tidlig diagnostik og dermed forbedre overlevelsen.

04 Background

Lung cancer

Epidemiology

Lung cancer (LC) has the highest incidence among cancers globally with almost 2.5 million new cases in 2022 accounting for 12.4% of all malignancies. It is also the leading cause of cancer-related deaths worldwide with 1.8 million deaths in 2020 (18%) [5–7]. Denmark holds the unfortunate distinction of having the 10th highest LC rate worldwide in 2020, counting both men and women, and the second highest rate in women, only exceeded by Hungary [8]. Among the Nordic countries, Denmark is by far ahead in terms of both LC incidence and mortality [9]. Despite an increase in survival over the last decade, five-year survival remains poor at 27% [10].

Diagnosis and treatment

LC is divided into four stages. Stages I and II are localized and often curable with surgery or radiation therapy, sometimes combined with adjuvant chemotherapy. Stages III and IV include metastases beyond the lung, i.e. to nearby lymph nodes (stage III) and distant organs such as the brain, liver, or bones (stage IV). At the later stages the cancer cells are more widespread, which makes complete surgical removal impossible. The patients diagnosed at a late-stage disease (60%) have poor prognosis [11]. Even with systemic treatment such as chemotherapy, immunotherapy, and targeted treatment, the disease is often too advanced to be cured. However, the introduction of checkpoint inhibitors targeting PD-1 and PD-L1, along with targeted treatments for EGFR mutations and ALK rearrangements, has extended survival and, in some cases, provided long-term disease control [12].

LC classification, mutation patterns and development

LC falls into two categories: Non-Small Cell LC (NSCLC), which accounts for 85% of cases, and Small Cell LC (SCLC), which makes up the remaining 15%. NSCLC is further divided into adenocarcinoma and squamous cell carcinoma, the two main subtypes. Different histological subtypes of LC are associated with distinct mutation patterns and

smoking history. **Figure 1** illustrates these associations, highlighting the most common subtypes and mutations where they are most prevalent. SCLC and squamous cell carcinoma are strongly associated with smoking, whereas adenocarcinoma is the most common subtype in non-smokers [13].

Smoking causes significant DNA damage through carcinogen exposure, resulting in a high mutational burden. Key mutations associated with smoking include KRAS, commonly found in adenocarcinomas, and TP53, which is prevalent across various LC subtypes. Non-smokers typically have a lower mutational burden, but their cancers are often driven by specific, targetable mutations such as EGFR and ALK, which are less common in smokers.

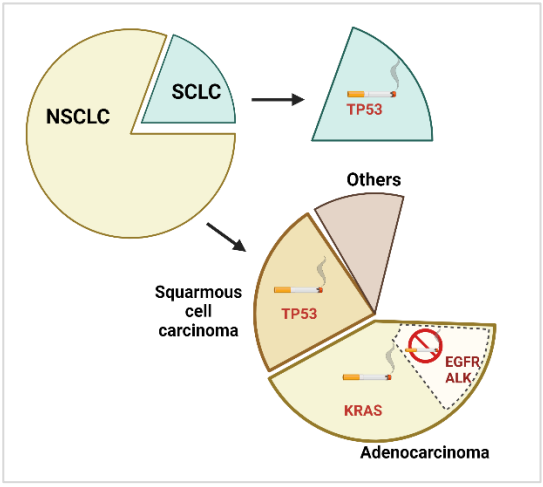


Figure 1: A simplified illustration showing the relationship between lung cancer histological subtypes, mutational patterns, and their association with smoking in the most common subtypes. Only the most frequently occurring mutations are displayed and noted for the subgroups in which they are significantly overrepresented. NSCLC: Non-small cell lung cancer, SCLC: Small cell lung cancer. Created using Biorender.com.

The development of LC is driven by the accumulation of genetic mutations over time. These mutations eventually lead to driver mutations that promote tumor growth, division, and resistance to cell death. Smokers, who have a higher mutational burden, face a greater risk of LC. Typically, LC develops over 10-30 years following prolonged exposure to carcinogens like tobacco smoke, although the timeline can vary based on the intensity and duration of exposure. Similar latency periods are seen with environmental carcinogens. In non-smokers with specific genetic mutations, e.g. EGFR and ALK, LC may still develop over several years. SCLC generally develops more quickly [14, 15].

LC risk factors - causality and associations

Smoking is the leading risk factor for LC, accounting for approximately 85% of cases. The best approach for preventing LC progression in current smokers is smoking cessation. Former smokers, however, continue to face an elevated risk of LC. To reduce LC mortality following cessation, screening with low-dose CT (LDCT) is the next best approach. Combining smoking cessation with LDCT screening is more cost-effective than using either method alone, especially when cessation takes place at younger ages [16]. Old age is the second most significant risk factor, as it involves an accumulation of mutations, prolonged exposure to risk factors, and reduced immune system function [17]. For non-smokers, other risk factors include radon exposure, air pollution, genetic predispositions such as ALK and EGFR mutations, second-hand smoke, and occupational hazards [18]. These factors are relevant but often difficult to obtain for research purposes.

LC often develops asymptomatically or mimics benign conditions, making symptoms usually noticeable only when tumors are advanced. The most common symptoms are cough, dyspnea, and haemoptysis, occurring in 65%, 50%, and 20% of patients, respectively. In general practice, the positive predictive value of these symptoms ranges from 1 to 5%, meaning that less than 5% of individuals with these symptoms are actually diagnosed with LC. Many cases are attributed to other conditions such as COPD and pneumonia. This non-specificity makes the symptoms less useful for LC screening purposes [19].

Several comorbidities, including COPD, cardiovascular disease, and diabetes, are associated with LC due to their strong links with smoking, aging, and systemic inflammation [20]. While general practice data on these conditions are not easily accessible, prescription medication records can serve as proxies for certain comorbidities [21]. Regular laboratory analyses have shown a relatively weak association with LC, primarily related to elevated inflammatory markers, low sodium levels in SCLC, and elevated measures due to metastasis involvement [22–25].

With this said, many studies focus on association between single risk factors and LC without accounting for the complexity of interactions between multiple risk factors and the potential for nonlinear relationships. Advanced statistical methods such as machine learning can be used to address these complexities and will be further discussed in the next sections.

LC diagnostics in Denmark

Since 2009, Denmark has had an LC fast-track pathway to address the poor survival rate compared to many other western countries. The pathway aims to reduce the delay to

diagnosis [26]. Patients are referred from general practice to the LC fast-track clinic based on a "reasonable suspicion" indicated by an X-ray or alarm symptoms such as hemoptysis or a persistent cough exceeding 4 weeks [27]. General practitioners can also refer patients directly for a computed tomography (CT) scan if they present non-specific symptoms and signs of cancer [28]. If CT scan findings necessitate further investigation, the LC fast-track clinics either initiate nodule follow-up or proceed with diagnostic procedures such as a Positron Emission Tomography (PET) scan, bronchoscopies or transthoracic needle biopsies [27]. The fast-track units have contributed to shorter time to diagnosis along with improved treatment options and smoking cessation measures, leading to better prognoses over the last decade [29]. However, individuals referred to the units are typically high-risk patients presenting late-stage alarm symptoms [11]. To diagnose patients at earlier stages before symptoms arise, it is essential to assess a broader at-risk population, making LC screening vital.

LC screening

LC screening has received considerable attention following the publiced results of the National Lung Screening Trial (NLST) and the Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON) trial. The NLST, conducted in the United States with over 53,000 participants, randomized individuals to receive either three annual screenings with LDCT or chest X-rays. The follow-up in 2011 after 6.5 years revealed a 20% reduction in LC mortality in the LDCT screened group [30]. The NELSON trial, conducted in the Netherlands and Belgium with over 15,000 participants, randomized individuals to receive LDCT screenings at regular intervals or standard care. The 10-year follow-up in 2020 demonstrated a 24% reduction in LC mortality for men and an up to 33% reduction for women [31]. While especially the NLST trial was initially criticized for the high rate of false-positives (around 25% in the first two rounds), the rate decreased over time, and the net benefit of a mortality reduction outweighed the risks associated with the false positives[32].

Several countries have either conducted or are conducting LC screening trials, including the UK (UKLS) [33], Italy (ITALUNG) [34], Canada [35], Brazil [36], France [37], Germany [38], Norway [39], Japan [40], and several in China [41, 42]. While most trials have been adequately powered, some were unpowered, among these the Danish LC screening trial (DLCST) which found no statistically significant difference in LC mortality between the two groups [43]. **Figure 2** depicts the relative risk of LC mortality in nine randomized controlled trials.

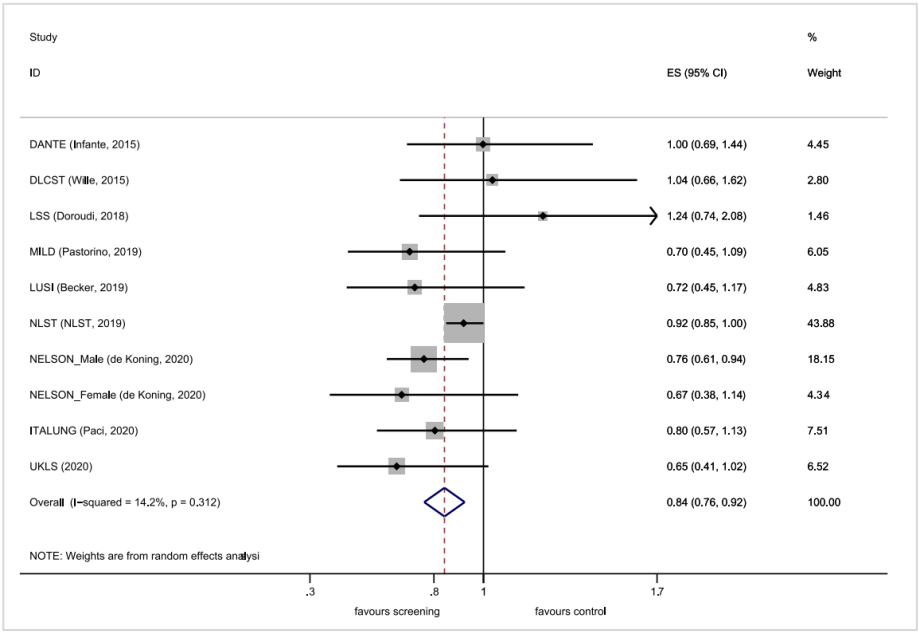


Figure 2: Forest plot showing the relative risk (RR) of lung cancer (LC) mortality across nine randomized controlled trials. The overall RR is 0.84 (95% CI: 0.76-0.92), indicating a 16% relative reduction in LC mortality among individuals who were screened compared to those in the control groups (represented by the dashed line). Source [33].

As a result of the cumulative evidence from numerous trials, the list of countries implementing additional follow-up pilots or regional LC screening programs is growing [44]. A recent initiative is the EU funded LC screening trial named 4-IN-THE-LUNG-RUN (Towards INdividually tailored INvitations, screening INtervals, and INtegrated co-morbidity reducing strategies in LC screening), in which five countries (the Netherlands, Germany, France, Italy and Spain) collaborate to develop and implement optimized, personalized LC screening programs for high-risk populations. They investigate recruitment methods and screening intervals and also focus on detecting coronary heart disease and chronic obstructive pulmonary disease (COPD) on LDCT scans. Within a period of two the study aims to include 26,000 individuals and follow them for at least five years [45].

Despite numerous economic evaluations finding LDCT cost-effective, the only nationwide implementation of LC screening is currently in the US. Their Preventive Services Task Force (USPSTF) recommends LDCT for individuals aged 50-80 who are current smokers

or have quit within the past 15 years, with a smoking habit of 20 pack-years [46, 47]. In Europe, the commitment to large-scale implementation has been cautious, and only Poland, Croatia, Italy, and Romania have formally committed to setting up nationwide LC screening programs [48]. Results of the 4-IN-THE-LUNG-RUN study is expected to pave the way for future implementation strategies in Europe.

Across varying demographics and health infrastructures, many countries encounter similar challenges, including selecting high-risk individuals, optimizing risk stratification strategies, and addressing recruitment and adherence issues [48–50]. These three themes will be discussed in the following subsections.

LC screening criteria

In the larger screening programs and pilots, the target population has been defined based on age and smoking intensity measured in pack-years. In the NLST study the age of the participants was 55-74 years compared to 50-71 years in the NELSON study. The enrollment criteria in NLST included current and former smokers who had quit within the past 15 years and had a smoking history of at least 30 pack-years [51]. NELSON included current and former smokers who had quit within the past 10 years and who had smoked ≥15 cigarettes/day for ≥25 years or ≥10 cigarettes/day for ≥30 years. Both studies collected detailed smoking histories using comprehensive screening questionnaires [30, 52].

While these criteria help identify a significant proportion of high-risk individuals, many LC cases in lower-risk groups or those with different risk profiles would be missed. For instance, adhering to the current USPSTF screening criteria would only detect 68% of LC patients in the United States [53]. Roe et al. demonstrated that only 46% of participants in the Danish LC Screening Trial would qualify for screening based on the NLST criteria, even though the Danish cohort consists of a selected group of heavy smokers [54]. Additionally, up to 20% of LC patients are never-smokers and would not be eligible for screening based on these criteria [55]. Studies have also indicated that the USPSTF 2013 criteria performed poorly in detecting LC among African-Americans and women, compared to whites and men [56–58]. These limitations underscore the need for ongoing research to refine screening criteria and develop more comprehensive risk assessment models.

Several studies are exploring the role of liquid biomarkers, including autoantibodies, microRNA, blood proteins and circulating tumor DNA for potential LC detection and risk stratification [59]. Other research focuses on exhaled volatile organic compounds as non-invasive biomarkers [60]. Additionally, some studies are developing risk models that combine blood-based biomarkers, imaging biomarkers, and clinical data [61]. Although some biomarkers are commercially available, many still require further validation.

Typically, these biomarkers perform well during the design phase when tested at the time of diagnosis but fail when tested in pre-diagnostic samples or in early-stage disease settings, likely because they are tested in controls outside the target population [62]. There is a need for further validation studies and randomized trials to evaluate the utility of biomarkers, which is both time and resource-intensive [59].

In this thesis, we have investigated the use of data obtained from electronic health records (EHRs) and clinical registers. This approach can be more cost-effective than extensive biomarker testing and can reduce unnecessary tests and procedures by targeting individuals at risk based on their medical history and other health data. Additionally, EHR data provide longitudinal information, tracking patients over time and allowing for monitoring changes in symptoms or laboratory results [63, 64]. Denmark benefits from a high degree of completeness in registry data, making these "dry" data valuable for research purposes in a Danish context [65].

Risk stratification models

The current criteria used in both NELSON and NLST dichotomize individuals based on age and smoking history, but research indicates that models considering the individual risk yield higher sensitivities [66]. Among the externally validated risk models are the Bach model [67], LC Risk Assessment Tool (LCRAT) [68], Liverpool Lung Project (LLP) model [69], and the Prostate, Lung, Colorectal, and Ovarian (PLCO) model PLCOm2012 [70]. Furthermore, the Helseundersøkelsen i Nord-Trøndelag (HUNT) model developed on a Norwegian population has been externally validated on the Danish LC Screening Trial population, in which is demonstrated superior performance to the NLST and NELSON criteria, and on a recent prospective cohort in Norway [54, 71].

The maturity and superior performance of the risk models over dichotomized selection criteria have led to their inclusion in National Comprehensive Cancer Network (NCCN) guidelines [72]. Comparison of the models, however, can be challenging due to differences in prediction periods and included risk factors. For instance, the Bach model predicts a 1-year LC risk, while LCRAT and LLP predict the 5-year risk and PLCOm2012 and HUNT the 6-year risk. Additionally, risk factors vary with some models incorporating COPD, family history of LC, and exposure to asbestos.

The models were developed using Cox proportional hazard ratios or logistic regression analyses [50]. While these conventional statistical methods are interpretable and ideal for some data structures, AI prediction models offer other advantages. AI models are capable of handling complex, high-dimensional and non-linear data, which aligns with the need for

increased usage of electronic health record (EHR) data in predictive modeling [73]. The following sections provide a brief overview of the AI-methods used in this thesis.

AI-based prediction models

ML is a subclass of AI, which enables algorithms to learn and make predictions based on data without being explicitly programmed. Examples of ML are decision trees, support vector machines, and Bayesian networks (BNs). Deep learning is a subclass of ML including more complex models organized in neural networks [73]. **Figure 3** depicts the map and subsets of these fields.

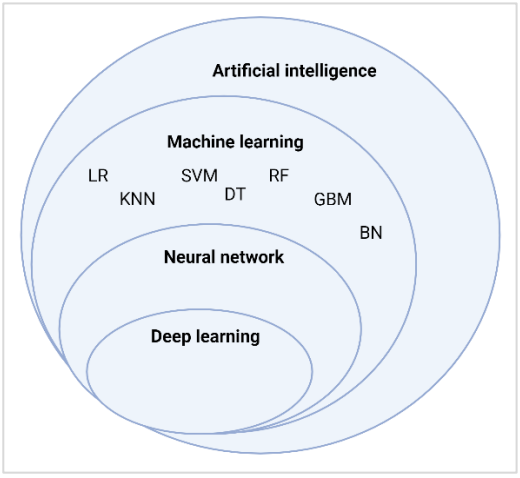


Figure 3: Venn diagram illustrating the structure of fields within Artificial Intelligence (AI). Machine learning (ML) encompasses techniques that learn from data without being explicitly programmed, including algorithms such as logistic regression (LR), K-nearest networks (KNN) support vector machines (SVM), decision trees (DT), random forest (RF), gradient boosting machines (GBM) and Bayesian networks (BNs). Neural networks are structures designed to mimic the human brain with multiple interconnected layers. Deep learning is a subset of neural networks involving multiple layers of neural networks. Created with Biorender.com.

Regardless of the AI-model used, the overall concept can be divided into three main categories; supervised learning, unsupervised learning and reinforcement learning (**Figure 4**). Supervised learning involves working with data where the outcome of interest (e.g. LC) is labeled or known and allows the algorithms to capture relationships between the label and the input variables. Depending on the outcome of interest, the problem can be classification or regression. Since LC is a binary outcome, this is a classification problem, while predicting a continuous outcome such as the length of stay in hospital would be a

regression problem. Unsupervised learning is based on unlabeled data and used to discover new patterns and structures in data, without knowing the status of the target variable in advance. This strategy has been used in the exploration of oncogenes in LC among genetic expressions [74]. Reinforcement learning is a third concept relying on feedback from the dataset, and the algorithm adjusts and improves based on rewards or punishments. This self-improving strategy has been used for nodule detection among other tasks [75].

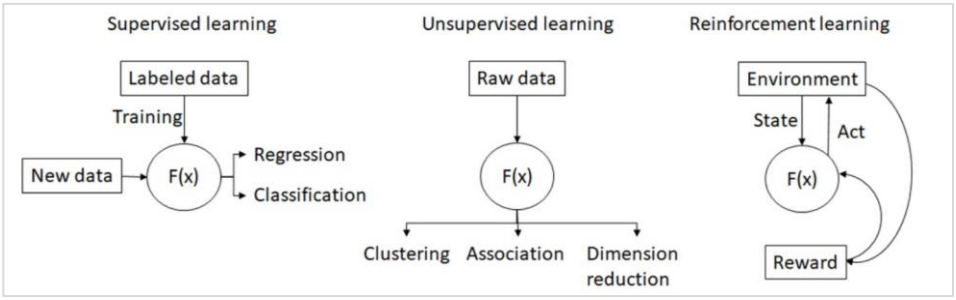


Figure 4: The overall concept behind supervised, unsupervised and reinforcement learning. Source [73].

ML prediction model pipeline

The general concept included in an ML pipeline, although many versions exist, is depicted in **Figure 5** and includes the below steps. It should be noted, however, that BN's follow a different framework.

1. **Data Collection and Cleaning**: In this phase, the dataset is gathered and properly labeled, e.g. identifying patients with LC. Correct labeling of the dataset is essential, as all subsequent steps rely on it. This idea is summarized in the well-known concept "garbage in, garbage out," meaning that poor-quality input data will lead to poor-quality output.

2. **Data Preprocessing**: This involves data exploration, where variable distributions are visualized and summary statistics are calculated to gain a thorough understanding of the dataset. It also involves handling missing values, encoding categorical variables, normalizing or standardizing variables, and potentially removing outliers.

3. **Dataset Splitting**: The dataset is mainly divided into a training set and a smaller validation set (sometimes referred to as a test set or hold-out cohort). The training data are used to train the different models, while the validation set is set aside for final evaluation. In addition, cross-validation is widely used for advanced model evaluation, since it ensures general performance across the entire dataset as opposed to just the results of a single test-

train split. In 5-fold cross-validation, the training set is divided into five folds. Each fold is used once as the test set (with masked LC status) while the remaining folds form the training set. This process is repeated (iterated) five times, shifting the test set with each iteration, and performance measures are averaged.

4. **Variable Selection**: This step reduces the number of variables to potentially enhance model performance by focusing on relevant variables. Techniques like principal component analysis (PCA) reduce variables into new ones that capture the highest degree of variance, optimizing results while maintaining simplicity and enhancing explainability. Other methods include statistical variable selection through e.g. a correlation coefficient or lasso regression. Variable selection based on SHAP (Shapley Additive exPlanations) values is another method described in later sections.

5. **Model Selection**: The goal is to identify the optimal model(s) for the specific problem, such as decision trees, logistic regression, or support vector machines.

6. **Model Training**: The selected models are trained on the training data with known status of the target variable (e.g. known LC status). After training, model calibration is often performed to adjust the predicted probabilities so they accurately reflect the true likelihood of the outcome.

7. **Model Evaluation**: Model performance is typically evaluated both on the test fold in cross-validation and on the hold-out validation set. A model that performs exceptionally well on the training data but poorly on the test or validation data may be overfitted. This implies that the model captures noise and random variation from the training data to the point where it fails to generalize to new data.

8. **Hyperparameter Tuning**: This involves fine-tuning the model to improve performance, such as optimizing the number of trees in a random forest. The model is then reassessed to ensure enhanced performance. This step can be done as an extra loop after measuring the performance, or as preparation before training the model [76].

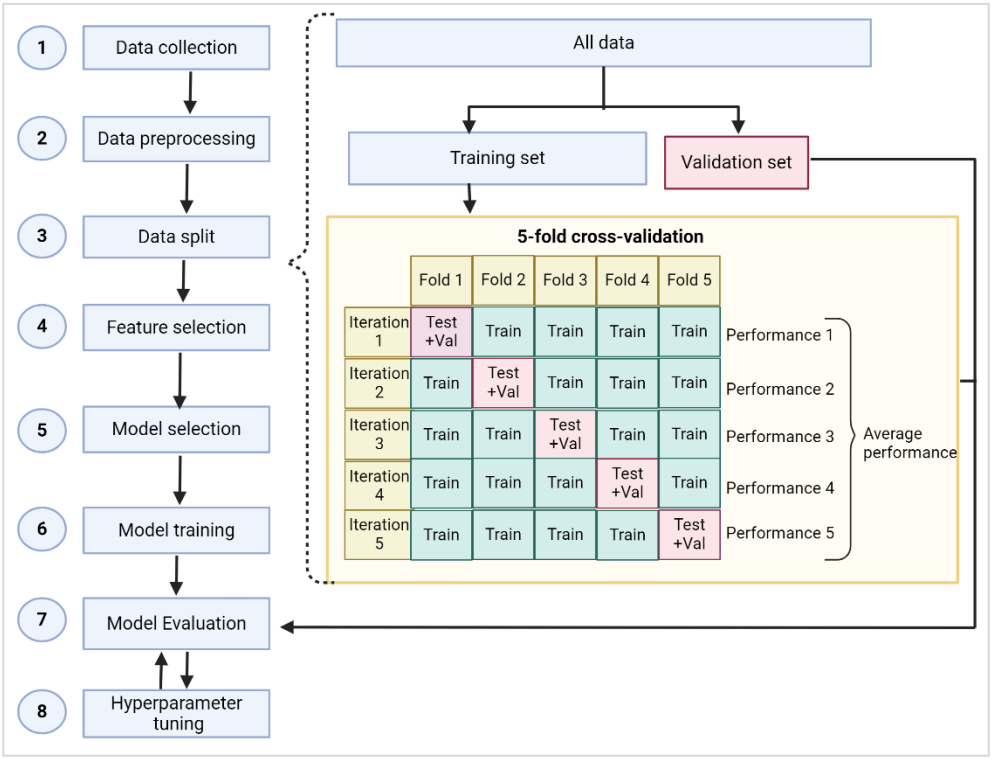


Figure 5: The overall ML pipeline and its involved steps. On the right, the data splitting process is detailed, showing how the data are divided into training and validation sets, and how the training set is used for cross-validation. Created with Biorender.com.

ML model selection

During the last decade the amount of available algorithms has increased exponentially in line with the increasing amount of available data and computational power. Even though model complexity is still increasing, some models remain widely used, while for instance decision trees continue to be built upon more complex frameworks such as the Random Forest, GBM or XGBoost. **Figure 6** gives an overview and simplified definition of the models commonly used in a supervised learning classification task such as detection of LC. The most common options for treating missing data are listed, and models are ranged by increasing complexity.

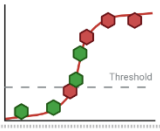
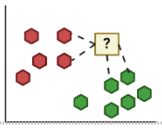
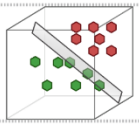
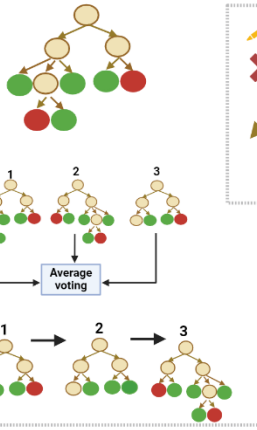
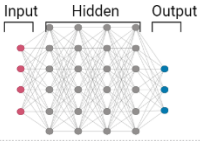
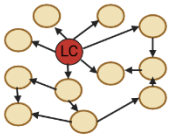
Definition	Illustration	Missing data	Complexity
Logistic regression • Linear model based on log-odds of target variable • follows sigmoid function		Imputation Deletion	
K-Nearest Neighbors • Non-linear relationship • Classification based on the majority class of the "k"- nearest neighbours		Imputation Deletion	
Support Vector Machines • Linear/non-linear • Finds the hyperplane/shape that best reparates data into different classes		Imputation Deletion	
Decision trees • Non-linear • Tree-like structure • Each branch represents a variable **Random Forest** • Non-linear • Combination of multiple decision trees • Each tree trained on different subset of the data • Averaged results based on majority voting **Gradient Boosting Machines (GBM)** • Non-linear • Sequential combination of multiple decision trees • Each tree corrects the errors of the previous on • XG-Boost is an optimized version of GBM		Imputation Deletion Create seperate branches	
Neural Networks • Non-linear • Network of artificial neurons • "input layer", "hidden layer", "output layer"		Imputation Deletion Various others	
Bayesian Networks • Non-linear • Probablistic graphical models • Baed on directed acyclic graphs (DAGs) • Each variable represented by a node		Predictions based on data distributions	

Figure 6: The most common machine learning (ML) algorithms used for classification tasks arranged by increasing complexity. Each algorithm is described in terms of its overall definition, options for handling missing data, and complexity. Created with Biorender.com inspired by [77–83].

Model explainability

The still more complex models create a need for explainability modules guiding the clinicians when advised by a certain prediction. SHAP is a widely used method that can be applied to most ML algorithms and models except BNs [84]. It is based on cooperative game theory and evolves around the Shapley values. By calculating SHAP values for each variable in a prediction, we can learn how this variable influences the model's performance. The most common visualization tools are the SHAP summary plot (**Figure 7**) and the force plot (**Figure 8**) [85]. The SHAP summary plot lists variables sorted by their importance. Each dot on the x-axis represents an individual's SHAP-value printed in red and blue for high and low variable values, respectively. The clustering of dots represents a frequently occurring SHAP value. The SHAP force plot gives insight into individual predictions and how individual variables contribute to the predicted risk. Each variable is represented by an arrow and the length of the error indicates the magnitude of the contribution. The direction shows whether it increases or decreases the predicted risk. The base value is the starting point, or the average prediction for all instances, and the final prediction or output value is the endpoint after considering all the variable contributions. Methods such as SHAP allow for usage in clinical practice and are possible solutions to the "black box" problem, which AI-models have been known as for years [86].

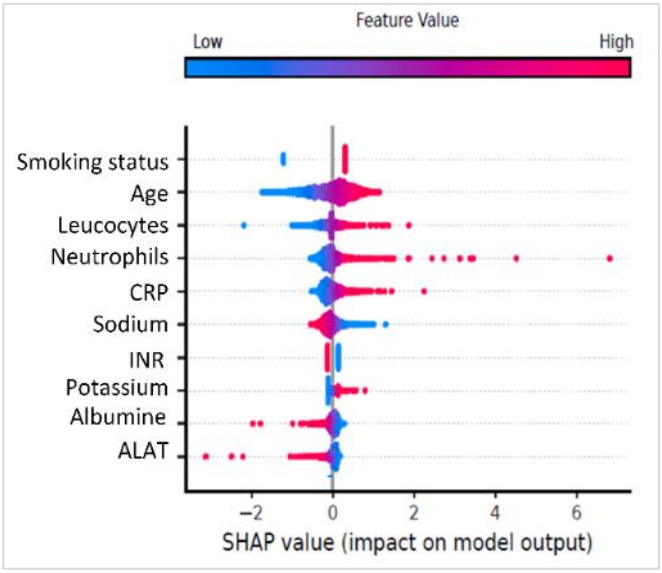


Figure 7: Artificially generated SHAP summary plot illustrating the most important variables, in this case in the prediction of LC. Variables are presented in descending order of significance. Each dot represents an individual SHAP value, with higher values indicating a greater impact on the model's output. Red dots represent high variable values, blue dots low values. In this plot, the most important variable is smoking status. A high value (indicating an active or former smoker) corresponds to a high SHAP value, thereby having substantial impact on the model. With inspiration from [2].

Figure 8: SHAP force plot showing the prediction of an individual's LC risk. This individual has a baseline risk of 0.46, which represents the risk before considering the specific variable values. After accounting for the variables, the predicted risk is 0.48. Given a default risk threshold of 0.5, a prediction of 0.48 indicates a non-LC patient. Variables in blue decrease the predicted risk, while variables in red increase it. For instance, the levels of LDH, creatinine, and calcium are the main factors reducing the risk prediction, whereas high age is the primary factor increasing the risk. Source [2].

BNs on the other hand, explain the relationship between variables by means of directed acyclic graphs (DAGs). In a DAG each variable is represented by a node, and the arrow indicates the direction of influence from one variable to another. The term “acyclic” refers to the one direction throughout the graph. Similarly to the SHAP force plot it is possible to see how the individual risk changes in a simple BN based on available data on the particular instance. The most well-known example is the BN “Asia network”, designed for educational purposes (**Figure 9**). It provides the probability of a patient having tuberculosis, LC or bronchitis based on conditional probabilities from various factors such as a visit to Asia, smoking habit, dyspnea, or a positive X-ray. **Figure 9A** displays the DAG representing the interconnection between variables. **Figure 9B** displays the same model visualized with probability distribution bars. It gives the probability distribution of the variable based on the network's initial state. Upon entering evidence, e.g. the absence of bronchitis, the probability distributions are updated to reflect the posterior probabilities given the new evidence (not shown in the figure). Overall, the DAG and belief bars provide insight into the probabilistic relationships within the BN. The DAG structures the dependencies, and the belief bars visually communicate the dynamic probabilities as evidence is incorporated [87, 88].

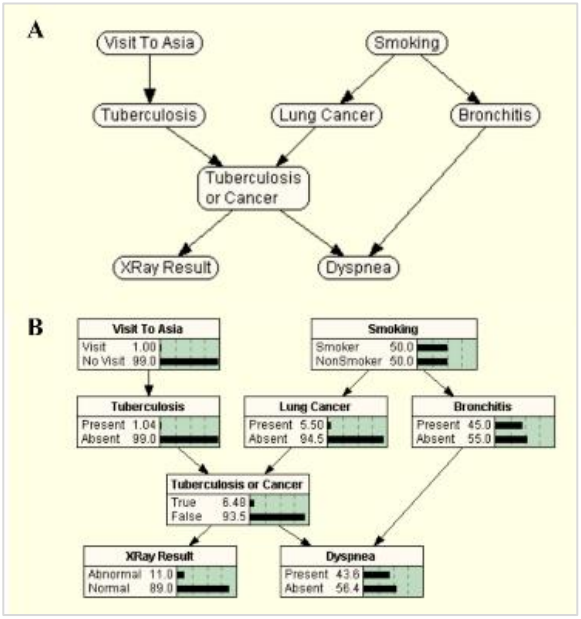


Figure 9: The Bayesian Network (BN) example known as the Asia Network. **A:** Directed acyclic graph (DAG) displaying connections between the nodes or variables. **B:** DAG transformed into a BN with probability distributions displayed as bars based on its initial state. Given evidence on a variable, e.g. the absence or presence of bronchitis, the network adapts and the probability of other variables change accordingly (not shown). Source [89, 90].

Prediction model evaluation and validation

Discrimination

Regardless of the prediction model chosen, evaluation metrics remain consistent. Assessing the model's performance involves evaluating its ability to accurately differentiate between true positives (e.g. LC patients) and true negatives (e.g. non-LC patients). This relationship is known as discrimination and visually represented on the ROC curve, with the true positive rate (TPR) plotted against 1 -true negative rate (TNR) on the x-axis. The measure of discrimination, the AUC, quantifies this performance. An optimal model achieves an AUC of 100%, forming a curve that hugs the top-left corner of the plot. Conversely, a model equivalent to random chance has an AUC of 50%, resulting in a diagonal curve. While the AUC provides an overall assessment of the model's discriminative ability, sensitivity (same as TPR) and specificity (same as TNR) can be assessed at different risk thresholds.

Figure 10 illustrates the ROC curve's foundation using a simplified example involving 10 patients, predicting LC status based on age **(A)**. A logistic regression model is applied to the data, assigning each individual a probability of LC **(B)**. To evaluate the model's performance, various risk thresholds, e.g. 0.9, 0.5, and 0.1 LC risk, are tested **(C)**. At each threshold, a 2x2 contingency table is generated displaying metrics such as TPR, TNR, false positive rate (FPR), and false negative rate (FNR). Plotting TPR against FPR on the ROC curve **(D)** maps each threshold to a specific point on the curve. For instance, setting a risk threshold of 0.1 or 10% requires classifying individuals with a risk probability exceeding 10% as LC patients, resulting in an 80% TPR and 40% FPR. By fitting a line across all thresholds the ROC curve is generated and the AUC can be computed. Although the AUC remains constant, sensitivity (TPR or recall), specificity (1-FPR), and positive predictive value (PPV or precision) fluctuate depending on the risk threshold.

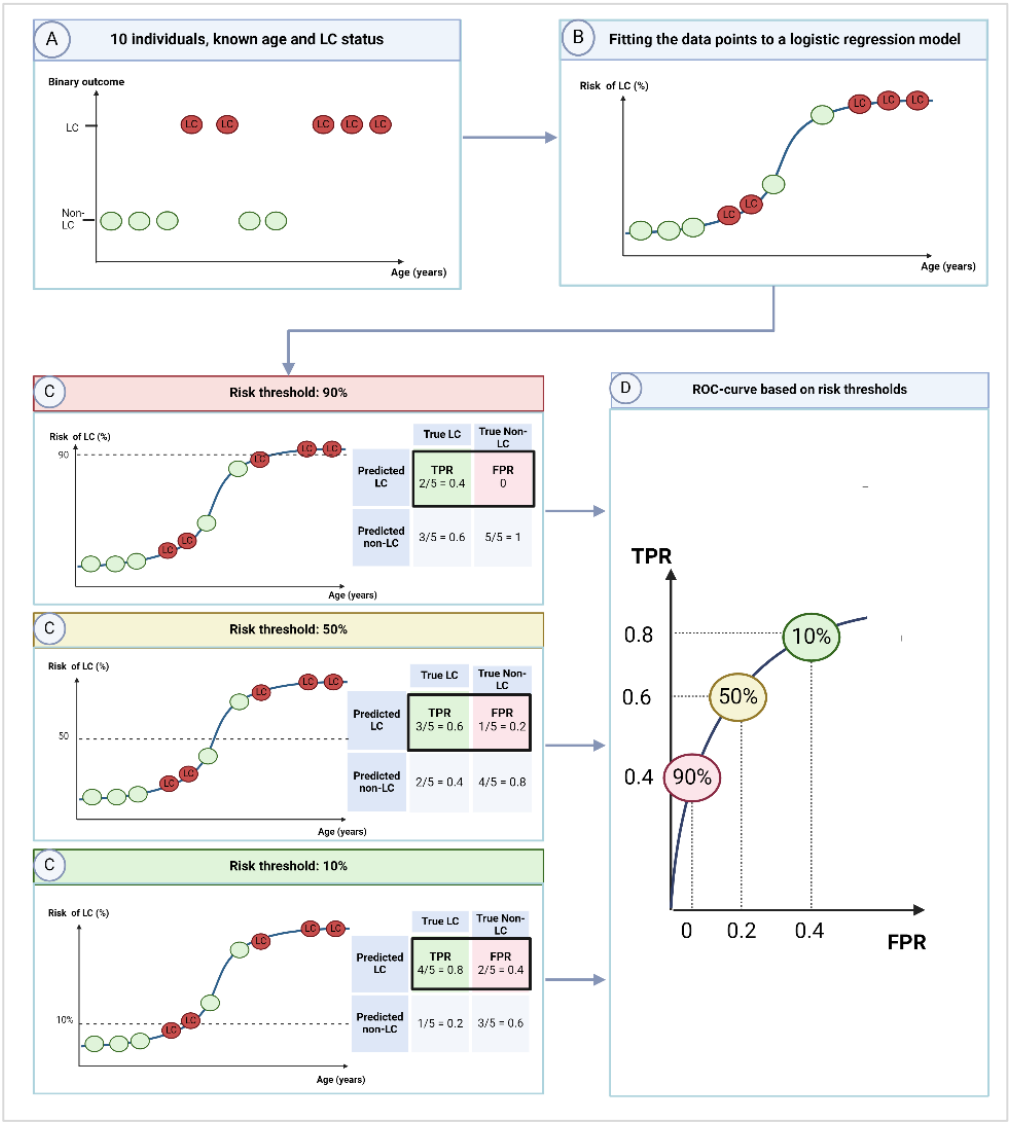


Figure 10: The foundation of the receiver operating characteristics curve (ROC) based on 10 patients with known LC status (5 LC depicted in red, 5 non-LC depicted in green) and age (A). When fitting the observations onto a logistic regression curve, the observations each get a probability or risk estimate of LC (B). C depicts the logistic regression curve when applying three randomly picked risk threshold (90%, 50% and 10% risk of LC). For each threshold the corresponding confusion matrix is depicted, holding the true positive rate (TPR) and false positive rate (FPR) in the first row. These

rates are plotted against each other on the ROC-curve, corresponding to a specific point on the curve (C). All possible thresholds are tested and the final points on the curve are connected with the best fitted line. The area under the curve (AUC) is a measure of overall discrimination at all thresholds. Created with Biorender.com.

Accuracy, another metric, reflects the proportion of correctly classified samples (both LC and non-LC cases) relative to all samples. However, this may not be ideal for rare diseases like LC, as high accuracy can be attained by correctly classifying all non-LC patients while misclassifying the sole LC patient. The F1-measure is a more balanced assessment by considering both precision (PPV) and recall (sensitivity), making it suitable for scenarios such as medical diagnosis where data are often imbalanced with rare outcomes [50, 91].

Calibration

While the metrics above assess the discrimination of a model, calibration is equally important to ensure that predicted probabilities align with actual probabilities. Calibration can be visualized using a box or curve plot showing actual probabilities against predicted probabilities. Discrepancies between predicted and actual probabilities, such as under or overestimation of the likelihood of LC within specific intervals, become evident from the calibration plot [91]. **Figure 11** displays an example of a calibration plot where the model underestimated the risk in lower-risk intervals while overestimating the risk at higher intervals. Calibration is especially important within the relevant decision risk-threshold, which is the cut-off level of risk where individuals are considered eligible for screening [92].

The Brier is a widely used measure to assess calibration calculated by the mean squared difference between predicted probabilities and actual outcomes, ranging from 0% for a perfect model to 0.25 for a non-informative model with a 50% risk of the outcome [91].

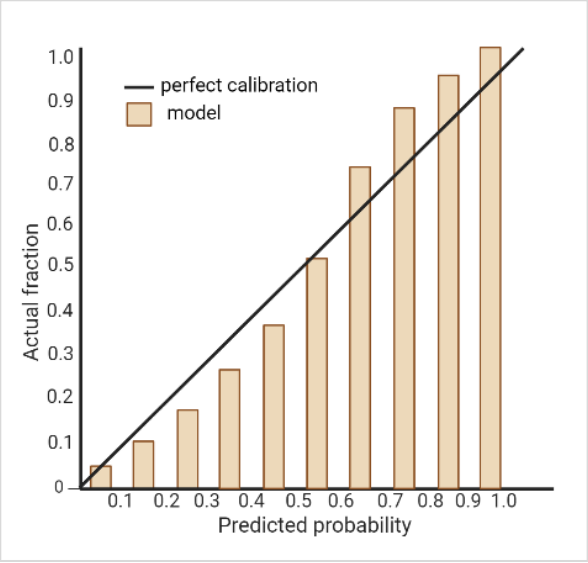


Figure 11: Calibration assessed using a predicted versus observed plot. The predicted LC risk is displayed on the x-axis, the observed risk on the y-axis. Both risks are represented in 10% interval bars. The diagonal line indicates the perfect calibration with predicted and actual risk aligned. Created with Biorender.com.

Clinical utility

The clinical utility of a prediction model can be evaluated by a decision curve plot in which the net benefit is plotted against threshold probabilities. The curve of a specific prediction model is compared to the scenario of treating all patients as well as treating no patients **(Figure 12)**. The model is of clinical utility if the net benefit is greater than the two extreme scenarios mentioned [91, 93].

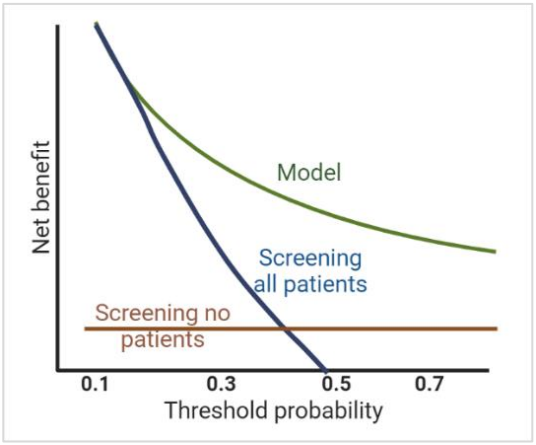


Figure 12: Decision curve plot illustrating clinical utility. The y-axis represents the net benefit and the x-axis the threshold probability. The specific model evaluated is depicted in green. For comparison, two extreme scenarios are included: screening all patients (blue) and screening no patients (brown). The model shows the same net benefit as flagging all patients up to a threshold of approximately 0.2. Above this risk threshold, the model demonstrates a positive clinical utility. The cross-section of the blue line and x-axis corresponds to the disease prevalence, in this example 50%. Created with Biorender.com.

Validation

While the aforementioned tests focus on evaluating a model within a specific population, the next step towards implementation involves further validation. Internally, a model is typically validated on a subset of the original cohort, such as data from the last year of the study period, to assess its robustness against the most recent data. Additionally, external validation is necessary, which may involve testing the model on a similar cohort from another region or hospital, or on a different cohort with a lower incidence of cases or a potentially different demography in regards to e.g. sex or ethnicity. Regardless of the setup, it is crucial to validate the model on a population resembling the target population in which the model might be deployed. This helps prevent disparities between performance in validation tests and future prospective trials. In the case of an LC detection or predictive model, external validation may include testing it on similar risk patients from a different region as well as on lower-risk populations such as those followed at the pulmonary clinic or in general practice. Testing the model across various, relevant populations ensures its robustness and evaluates its utility for implementation in different settings [50, 91].

Selecting the risk cut-off

When presenting the results of a prediction model, discrimination and calibration is often assessed through the AUC or Brier-score and compared to other risk-models on the same measures. Nevertheless, the risk cut-off for e.g. referral to an LDCT scan has to be set before a risk model continues to the step of prospective pilot studies. The relevant cut-off depends on the population and specifically on the purpose of the test. If a test is constructed

to “rule-out” a disease, e.g. if the test was potentially used to select relevant patients for a CT scan, it would have to be very sure not to miss any potential LC patients. Hence, it is relevant to aim for a cut-off that maximizes the sensitivity. In case the purpose of a test is to “rule-in” a disease, a cutoff that maximizes the specificity should be aimed for. While screening for LC seems like a clinical scenario aiming to maximize sensitivity in order not to miss any cases, screening scenarios are more often thought of as rule-out scenarios. Since most screening models do not have perfect discrimination, a maximized sensitivity would naturally come with a high FPR leading to an overflow of excess CT scans, possible adverse events and worried patients. Consequently, when finding the ultimate risk threshold for a screening model, the number-needed-to-screen also has to be taken into account. If lowering the threshold to e.g. 10%, more patients would be screened, leading to a higher number-needed-to-screen to encounter one LC patient. If working with a high threshold such as 90%, only the patients in the top 10% risk interval would be screened, leading to a lower number-needed-to-screen to find one LC patient [91]. To provide a context for these measures, the number-needed-to-screen for the detection of one LC was 92-133 per round in NELSON compared to 97-147 in the NLST [32, 94].

Suggesting an appropriate cut-off is challenging and requires thorough evaluation through cost-efficiency analyses that consider available healthcare resources and potential benefits and harms. However, the literature has proposed several potential cut-offs. For example, a threshold of 1.51% from the PLCOm2012 model has demonstrated improved performance compared to the NLST criteria [95, 96]. Based on several validation studies of risk models, the NCCN clinical practice guidelines now recommend using LC screening models to assess the risk for certain high-risk individuals who fall outside the NLST criteria, often due to younger age or additional risk factors. The NCCN suggests screening if the risk score exceeds 1.3% [72]. Until now, only a few trials have implemented the use of risk models in the selection of participants. Among them is the Ontario LC screening trial, where the PLCOm2012 model was used to select participants for LDCT based on a risk cut-off of 2% [97]. Additionally, the UKLS trial was the first randomized screening trial to select patients for LDCT based on individual risk scores, using a risk cut-off of 4.5% derived from the LLPv2 risk model [33]. In the current 4-IN-THE-LUNG-RUN trial, individuals are randomized to participate based on dichotomized age and smoking criteria or individual risk stratification, with a cut-off above 2.6%, derived from the PLCOm2012NoRace model [45].

AI from the patient’s perspective

While the advantages of using AI-based risk models for LC screening may be evident to clinicians, it is essential to consider patients' perspectives on AI. A 2017 European Union

survey conducted in homes across 28 countries, including Denmark, revealed that the Danish population is notably more knowledgeable about and supportive of AI compared to the EU average. Specifically, 82% of Danish participants held a generally positive view of AI, compared to 61% across Europe. Additionally, 87% of Danes would consent to sharing their health data, whereas the European average stands at 70% [98]. In contrast, a 2022 survey of 11,000 Americans found that 60% would be uncomfortable with their healthcare providers relying on AI [99, 100].

Multiple ethical considerations also arise with AI prediction models. Labelling individuals as "high-risk" for future conditions can induce stress and anxiety, especially if the event is years away. Even though the General Data Protection Regulation (GDPR) requires explicit consent for processing personal health data, including for AI-driven predictions, there is currently no established practice for obtaining patient consent specifically for AI-enhanced care. Consequently, many patients remain unaware of its background role [101]. While GDPR's rules are not inherently flawed, they may need adaptation to align with the evolving landscape of AI in medicine as the technology becomes more integrated and regulated [102]. GDPR also mandates the right of patients to understand predictions, including the logic behind them and their consequences. This necessitates AI systems to incorporate explainability modules such as the SHAP module to ensure transparency [103, 104].

Patient recruitment and adherence

While numerous LC prediction models have shown good results in discrimination and some have been successfully externally validated, the majority still rely on data difficult to obtain from the general population. Such data, e.g. smoking history, are typically collected through population questionnaires to which high-risk individuals often do not respond or from general practice [105]. The American Lung Association states that in 2022, only 5.8% of screening-eligible patients underwent screening [106]. Consequently, adherence to screening remains a significant challenge.

In Denmark, the nationwide digital mailbox system, "e-Box," is used by public authorities to reach all citizens (alternatively by post) [107]. The tool is widely used to distribute official information, but relying solely on responses via e-Box for screening can be problematic. In the NELSON study more than 606,000 individuals were approached of which only 150,920 (25%) responded and 15,822 consented (2.6% of all approached) [108].

To ensure optimal recruitment and adherence, screening needs to be easily accessible and fit into the everyday lives of typical high-risk patients. As one of the coordinators of the NLST stated, it is essential to "go where the smokers are" [51]. Pilot programs using a

screening bus have shown promising results, e.g. the Manchester Lung Health Check Pilot launched in the UK in 2016. It used the PLCOm2012 model with a 6-year LC risk threshold of 1.51%, resulting in 1,384 screened individuals, 3% of whom were diagnosed with LC and 80% of them at an early stage [109]. In the NLST, patients were recruited through nursing centers, mail, community outreach, and mass media [30]. Another effective approach is reaching patients at the caretaker level, e.g. in general practice or hospitals, where they come for regular check-ups [51].

COPD is a known independent risk factor of LC and patients with moderate to severe COPD usually attend annual checkups at pulmonary clinics [110, 111]. Consequently, investigating LC incidence and crossover to the LC-fast track clinic has been of interest in this thesis to further apply risk models to the population.

05 Objectives

The previous sections have emphasized the importance of early LC detection and detailed the current diagnostic framework, including LC fast-track clinics in Denmark. They also described the state of LC screening programs and the challenges related to selection criteria, stratification, and adherence. Additionally, AI-based prediction models were introduced, including BNs.

The primary objective of this thesis is to investigate AI-based LC detection models for patients at high risk of LC. Moreover, it aims to explore the potential of LC screening in a medium-risk cohort of COPD outpatients. The main focus is on data sourced from clinical health records and registries. **Figure 13** depicts the five studies and the relation between them.

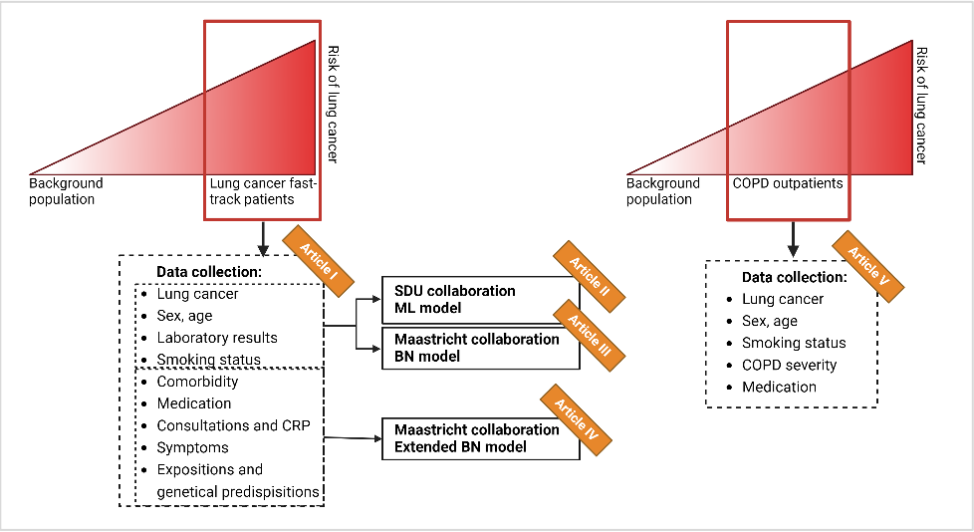


Figure 13: Outline of the thesis context and the interconnection of the five articles. SDU: University of Southern Denmark. COPD: Chronic obstructive pulmonary disease. BN: Bayesian Network. ML: Machine learning. Created with Biorender.com.

Article I seeks to analyze the risk factors and associations with LC in high-risk patients undergoing examination in fast-track LC clinics.

Articles II-IV draw upon data from the same cohort to develop LC detection models. Article II employs smoking and laboratory data to construct an ML model, while Article III

utilizes a BN model. The remaining dataset is employed in expanded BN models (Article IV) aimed at enhancing performance through enrichment of the dataset.

Article V aims to assess LC incidence and stage distribution among COPD outpatients, thereby evaluating the relevance of extending LC screening models to this demographic.

The next section will detail methods and results; articles I and V separately f and articles II, III, and IV combined due to similarities. Subsequently, a unified discussion will follow encompassing all articles. The ethics statement described under Article I covers all five studies.

06 Methods and results

Article I: Methods

AIM Article I: Analyze the risk factors and associations with LC in high-risk patients undergoing examination in fast-track LC clinics.

Study cohort

This was a retrospective observational study including all individuals in the region of Southern Denmark examined on suspicion of LC within the 10-year period of 1st of January 2009 to 31st of December 2018. They were identified using the classification codes AFB26 and Dz031b. The first code indicates initiation of examination in the LC fast-track clinic and the second code that the individual is tentatively under observation for LC. Both codes belong to the Danish health care classification system (SKS) [112] and data were extracted from the regional data warehouse. LC patients were identified based on the International classification of diseases (ICD), version 10, code C34 (malignant neoplasm of bronchus and lung) delivered from the Danish Lung Cancer Registry [113]. 1,646 individuals diagnosed with LC outside the LC fast-track pathways were also included, who were without any of the two SKS-codes applied. A total of 283 patients with previous LC (2002-2009) and 56 patients with missing information on sex were excluded. The final cohort consisted of 38,944 patients of whom 11,284 had LC (29%) and 27,660 (71%) did not **(Figure 14)**.

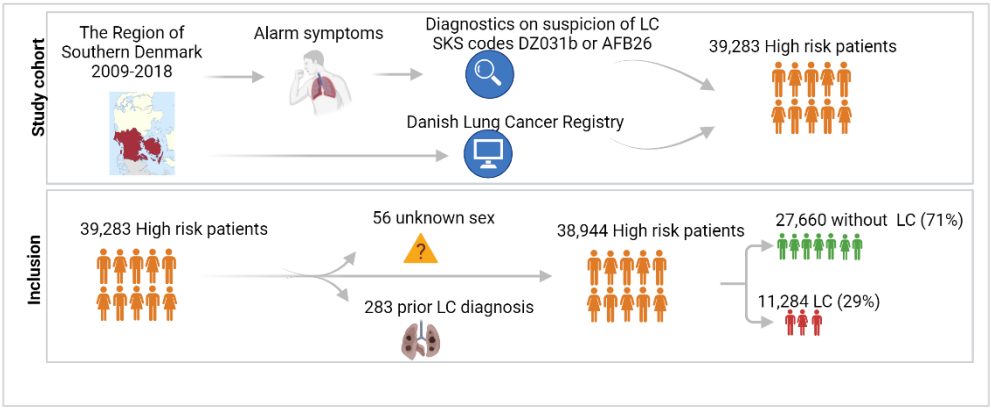


Figure 14: Study cohort and inclusion based on clinical coding (SKS-codes) and lung cancer (LC) registrations from the Danish Lung Cancer Registry. Created with Biorender.com.

Data sources and variables

The date of LC diagnosis was used as the reference point for LC patients and that of the first assigned SKS-code for non-LC patients. This was referred to as the "index-date". The regional data warehouse delivered all input data listed in **Table 1**. This table furthermore provides an overview of the data origin from structured codes or free text extracted from the EHRs, along with their time frames relative to the index date and data format. Additionally, it outlines the interpretation of missing data, distinguishing between actually missing observations and instances where the absence of a condition or test can be inferred.

Table 1: Overview of included data categories. MCAR (missing completely at random), ICD-10 (international classification of diseases, version 10), ATC (anatomic therapeutic chemical code), EHR (electronic health records). BCC and Labka are names of the current and former laboratory system used in the Region of Southern Denmark. MCAR: Missing completely at random. MAR: Missing at random. LC: Lung cancer.

Data category	Data origin	Time frame relative to index date	Format	Interpretation of missing data

Comorbidity	ICD-10 codes	2 years before	Binary (present/ absent)	Not missing but absence of disease
Medication	ATC-codes	6 months before	binary (prescribed/ not prescribed)	Not missing but absence of prescription
Smoking status	Free text from keywords in EHR	Unrestricted	Binary (never vs. Current/ former)	MCAR since this is an included part of the examination in the LC fast-track clinics / MAR for certain periods
Consultations in general practice	Registered services	6 months before	Continuous and categorical (cutoff 0 and 4 registrations)	Not missing but absence of consultation
C-reactive protein point-of-care tests in general practice	Registered services	6 months before	Continuous and categorical (cutoff 0 and 4 registrations)	Not missing but absence of test
Laboratory results	BCC and Labka, 20 relevant analyses	28days before, 2 weeks after	Continuous variables	MCAR since this is an included part of the examination in the LC fast-track clinics / MAR for certain analyses and periods
Symptoms	Free text from EHR, 20 selected symptoms	4 weeks before, 2 weeks after	Binary (present/ absent)	MCAR since this is an included part of the examination in the LC fast-track clinics/ MAR for certain periods

Familial predispositions	Free text from EHR, siblings/ parent with LC	4 weeks before, 2 weeks after	Binary (present/ absent)	MCAR since this is an included part of the examination in the LC fast-track clinics/ MAR for certain periods
Exposures	Free text from EHR, predefined exposures	4 weeks before, 2 weeks after	Binary (present/ absent)	MCAR since this is an included part of the examination in the LC fast-track clinics/ MAR for certain periods

Comorbidity data

Comorbidity data were collected as ICD-10, reflecting hospital encounters. ICD-10 codes from in the Charlson comorbidity index (CCI) were included with minor adaptations for any malignancy. A detailed description of included ICD-10 codes and weights for calculating the CCI-index can be found it Article I, p.4 and Tables S1 & S2 [1]. A 2-year period preceding the index date was chosen as a reference point, since diseases not recorded during this timeframe were deemed unlikely to have clinical relevance at the time of LC diagnosis. Data were converted into binary format, indicating either the absence or presence of a disease. The two SKS codes used to identify the population originated from the same dataset as comorbidity data. Given that all patients included had SKS codes, individuals without ICD-10 codes were considered as lacking the specific disease and therefore not classified as true missing data.

Medication data

Since symptoms of LC, pneumonia and COPD often overlap, we included the most common antibiotics used to treat pneumonia as well as corticosteroids and inhalation devices used to treat COPD. Different antidepressants were also included as a proxy for depression and anxiety. All medications were based on anatomic therapeutic chemical (ATC) codes. For a list of the underlying ATC codes, please refer to Article 1, supplementary material. A period of six months was considered relevant to identify medications in relation to the time of diagnosis. The data were converted into binary format, indicating the presence or absence of a specific medication. Individuals with no registrations were recorded as no drugs prescribed rather than classified as missing.

Smoking data

Data on smoking status were collected based on free text from the EHR subfields related to smoking and risk factors without any restriction in time. In case of multiple notes per patient, the note that provided the most comprehensive details on smoking status was selected. Patients with detailed information on pack-years were primarily categorized as active smokers and the remaining as active-smoker or former-smoker, or status unknown. To resolve the "unknown" category, additional notes for these patients were evaluated and a smoking label was assigned based on the note containing the most comprehensive information. Any duplicate entries were removed, retaining only the note responsible for the patient's label. Missing data on smoking status was considered to be MCAR (missing completely at random) since it was expected that all patients entering the LC fast-track units would be asked about smoking habit as part of the examination. Hence, missing information could be due to the clinician not asking about or documenting the smoking habit, or potentially registering smoking habit in a different subfield in the EHR. However, the implementation of the EHR was a gradual process during the first years of the study period, which could explain the lack of input during the first years. In that case missing data for the first study years could be considered to be missing at random (MAR).

Consultations and CRP point-of-care tests in general practice

In Denmark, general practitioners are paid through a combination of fee-for-service and capitation, funded by the healthcare region [114]. Each consultation and CRP point-of-care test is billed using a specific code and an exact fee, and all codes are centrally registered. The number of consultations and CRP point-of-care tests recorded within six months before the index date were included and analyzed both as continuous variables and categorical variables with a cut-off of zero and four registrations. Individuals without any registrations were noted as having an absence of registrations, and no data were considered missing.

Laboratory results

Laboratory results were collected from both the current system (BCC, 2011-2018) and the former system (LABKA, 2008-2011). The set of analyses used in the LC diagnostic unit at Vejle Hospital, which included 20 different laboratory tests, was used for further investigation. These tests were B-hemoglobin, P-sodium, P-potassium, P-lactate dehydrogenase (LDH), P-alanine transaminase (ALAT), P-CRP, P-creatinine, P-international normalized ratio (INR), P-total calcium, P-albumin, P-amylase, P-total bilirubin, P-alkaline phosphatase, and counts of B-basophils, B-neutrophils, B-leucocytes, B-monocytes, B-lymphocytes, B-eosinophils, and B-platelets. Laboratory results were

collected within the time frame of 28 days before the index date and 14 days after to capture the samples taken at the time of examination for suspected LC. Only samples ordered by the four diagnostic units in the region (Esbjerg, Odense, Sønderborg, and Vejle) were included. In case of multiple samples per patient, the sample with the highest number of included analyses was chosen for examination. Individuals with less than 17 analysis results were excluded. All results were analyzed as continuous variables.

Missing data were overall considered to be MCAR since all patients were expected to have laboratory samples taken around the time of examination. However, some analyses, e.g. amylase, INR, and bilirubin, were rarely taken at certain hospitals. In these cases, the absence of data could be attributed to systematic reasons. Therefore, one could argue for classification as MAR due to the clinical relation at the time and place.

Symptoms, familial predispositions and exposures

Free text from all EHR notes was collected within four weeks before and two weeks after the index date. The note deemed to belong to the LC fast-track system was selected for annotation. Two medical students and a medical doctor manually annotated the symptoms, i.e. hemoptysis, pneumonia, cough, dyspnea, fever, weight loss, fatigue, hot flashes, hoarseness, back pain, other pain, angina, headache, dizziness, and edema. Conditions referring to the examination period were marked as present and as absent if not.

Familial predisposition to LC was recorded as present if a reported sibling or parent had LC. Exposure was recorded as present if the individual had a history of working with radon, asbestos, nickel, chromium, aromatic hydrocarbons, or welding, regardless of duration.

All symptoms, familial predispositions, and exposures were analyzed as binary variables, indicating either the presence or absence of the condition. If no EHR note pertaining to the LC fast-track clinic was found, a mark of having missing information on symptoms, familial predispositions, and exposures was attached to that patient. Such missing data were considered MCAR since it was expected that all patients would be asked about these matters in the LC fast-track clinic. However, in line with registration of symptom data, the gradual implementation of the EHR system could lead to missing data over the first study years, and missing data from this period would consequently be considered MAR. If the note corresponded to the consultation in the LC fast-track clinic, absent symptoms, predispositions, and exposures were noted as absent, not missing.

Statistical analyses

The distribution of patient and data variables related to LC are presented as percentages for categorical variables and as median with interquartile range (IQR) for continuous variables. Associations between groups were examined using the Chi-squared test for categorical variables and the Wilcoxon signed-rank test for non-parametric continuous variables. All statistical tests were two-sided with a significance level set at $P<0.01$ to account for multiple testing. The proportion of missing data was independently assessed for each variable and evaluated to determine if it was missing at random or systematically, which could introduce selection bias. Subgroupanalyses compared associations between groups within the cohort with complete data. All statistical analyses were conducted using Stata version 17.0.

Ethics (covers all four studies)

All studies were conducted in accordance with the Declaration of Helsinki (as revised in 2013) and approved by the Region of Southern Denmark (19/30673, 06-12-2020) and the Danish Patient Safety Authority (3-3013-3132/1, 03-30-2020). Individual consent for these retrospective analysis was waived. In respect to patient autonomy and in accordance with GDPR, all data were pseudonymized and uploaded to a cloud service for sharing with collaborators. The Region of Southern Denmark approved the sharing of pseudonymized data, provided that the data were indecipherable and did not contain substantial personal details such as extensive free text. Collaborative and data processing agreements were signed with all involved parties.

The diverse and large-scale datasets utilized in these studies qualify as big data, making data security and ethical considerations particularly important. Prior to the commencement of the research, a Data Protection Impact Assessment (DPIA) was conducted, outlining potential risks, the AI analyses involved, and compliance with data protection regulations across all studies. In developing the models, patient safety was ensured, as the retrospective nature of the data posed no harm to individuals. This project encompassed regional data from several hospitals, thus representing different demographics.

The results of these studies potentially offer significant benefits, including improved patient outcomes and survival rates, reduced need for systemic treatments, optimized resource allocation, and an increase in life years gained. With advancements in AI and predictive modeling, neglecting to implement tools that facilitate earlier detection could deprive patients of these improved outcomes. Ethically, the existence of such technology brings a responsibility to leverage it in accordance with the principle of beneficence. Therefore, conducting these studies is not only scientifically important but also ethically aligned with the goal of improving patient care.

Article I: Results

Demographics and LC Incidence and Stage

During the study period 38,944 individuals were examined on suspicion of LC of which 27,660 were diagnosed with the malignancy (**Table 2**). Over the years, the number of individuals examined increased, largely due to the growth of the non-LC group, which rose from 61% of all examined patients in 2009 to 75% in 2018. Conversely, the proportion of LC patients declined from 39% in 2009 to 25% in 2018. LC patients were notably older than non-LC patients, with a median age of 70 years (IQR, 63-77) and 67 years (IQR, 56-75), respectively, ($p<0.001$). Additionally, the LC group had a higher proportion of females (48% vs. 45%, $p<0.001$). The proportion of patients diagnosed with LC at stage I-II increased over the study period from 18% in 2009 to 31% in 2018.

Table 2: Demography and LC stage, LC (Lung cancer).

Demography	**Non-LC**	**LC**	**P-value**
Total, no. (%)	**27,660 (100)**	**11,284 (100)**	
Age, median (IQR)	67 (56-75)	70 (63-67)	<0.001
Females, no. (%)	5,461 (48.4)	12.515 (45.3)	<0.001
LC stage, no. (%)			
I		2,001 (17.7)	
II		914 (8.1)	
III		2,242 (19.9)	
IV		5,440 (48.2)	
Unknown		687 (6.1)	

Comorbidity

Comorbidity based on the included ICD-10 codes was not registered with 45% of the cohort within the two-year interval. Other malignancies were more prevalent in the non-LC than the LC cohort (14.0% versus 11.7%, p<0.001) (**Table 3**). The most common other malignancies were colorectal and breast cancer, both significantly more frequent in the non-LC group than the LC group (2.3% versus 1.4% and 2.0% vs. 1.4%, respectively, p<0.001 for all). Chronic pulmonary disease and pneumonia were the second and third most common with no statistically difference between groups. A higher fraction of LC patients was diagnosed with cerebrovascular disease, peripheral vascular disease and metastatic solid tumor compared to the non-LC group (p<0.001 for all). The CCI was zero in 62% of the LC patients compared to 65% of the non-LC group.

Table 3: Comorbidities and other malignancies. LC (lung cancer). Data are displayed in numbers and percentages, the number of pneumonias and the sum charlson comorbidity index (CCI) which is accompanied by the median.

Comorbidity data	**Non-LC N=27,660**	**LC N=11,284**	**P-value**
Myocardial infarction	454 (1.6)	225 (2.0)	0.02
Congestive cardiac failure	198 (0.7)	69 (0.61)	0.26
Peripheral vascular disease	828 (3.0)	555 (4.9)	<0.001
Cerebrovascular disease	915 (3.3)	525 (4.7)	<0.001
Dementia	200 (0.7)	77 (0.7)	0.67
Chronic pulmonary disease	3,379 (12.2)	1,429 (12.7)	0.22
Rheumatic disease	533 (1.9)	228 (2.0)	0.54
Liver disease	198 (0.7)	65 (0.6)	0.13
Diabetes mellitus	1,245 (4.5)	566 (5.0)	0.03
Hemiplegia or paraplegia	30 (0.1)	18 (0.2)	0.19
Renal disease	522 (1.9)	179 (1.6)	0.05
Metastatic solid tumor	772 (2.8)	6.5 (5.4)	<0.001

AIDS/HIV infection	19 (0.1)	9 (0.1)	0.94
Pulmonary tuberculosis	48 (0.2)	8 (0.1)	0.02
Sarcoidosis	79 (0.3)	18 (0.2)	0.02
Interstitial lung disease	194 (0.7)	74 (0.7)	0.62
Abscess	157 (0.6)	43 (0.4)	0.02
Pleural disease	725 (2.6)	297 (2.6)	0.95
Any pneumonia	2,944 (10.6)	1,132 (10.0)	0.07
Number of pneumonia	0 (50)	0 (50)	0.06
Sum CCI	0 (50)	0 (50)	<0.001
Other malignancies			
Colorectal	625 (2.3)	162 (1.4)	<0.001
Breast	552 (2.0)	163 (1.4)	<0.001
Prostate	408 (1.5)	149 (1.3)	0.24
Lymphoma	298 (1.1)	57 (0.5)	<0.001
Leukemia	225 (0.8)	89 (0.8)	0.81
Head and neck	179 (0.7)	108 (1.0)	0.00
Melanoma	193 (0.7)	44 (0.4)	<0.001
Bladder	154 (0.6)	75 (0.7)	0.21
Kidney	163 (0.6)	37 (0.3)	0.001
Mesothelioma	188 (0.7)	9 (0.1)	<0.001
Unknown primary tumor	83 (0.3)	50 (0.4)	0.03
Brain	30 (0.1)	100 (0.9)	<0.001
Esophagus and stomach	56 (0.2)	37 (0.3)	0.02
Others	354 (1.3)	127 (1.1)	0.21
Any of the above	3,859 (14.0)	1,321 (11.7)	<0.001

Medication

In 27% of the study cohort no ATC code was recorded within the 6-month interval of the study period. Patients with registered ATC codes were generally older and included a higher proportion of LC patients compared to those without any registered ATC codes. The proportion of LC patients with any prescription was significantly higher than that of non-LC patients (77.4% versus 72.0%, p<0.001).

All drugs, except for antibiotics, were prescribed to a higher percentage of LC patients compared to non-LC patients (**Table 4**). The non-significant difference in antibiotic prescriptions aligned with the non-significant difference in the proportion of patients with registered pneumonias.

Table 4: Prescription medication

Prescription medication	**Non-LC no. (%)**	**LC no. (%)**	**P-value**
Total	**27,660 (100)**	**11,284 (100)**	
Antibiotics	12,130 (43.9)	4,954 (43.9)	0.93
COPD inhalators	7,028 (25.4)	3,490 (30.9)	<0.001
Beta blockers	5,091 (18.4)	2,339 (20.7)	<0.001
Calcium antagonists	4,771 (17.3)	2,421 (21.5)	<0.001
ACE inhibitors	3,317 (12.0)	1,730 (15.3)	<0.001
Glucocorticoids	2,770 (10.0)	1,534 (13.6)	<0.001
SSRI	2,315 (8.4)	1,038 (9.2)	0.01
Metformin	1,694 (6.1)	837 (7.4)	<0.001
Other antidepressants	1,640 (6.0)	776 (6.9)	<0.001
TCA	582 (2.1)	310 (2.8)	<0.001

Smoking status

Notes containing relevant information on smoking status were available for 23,006 patients (60% of the total cohort). As expected, the amount of missing data was highest in the early years of the study, likely due to the implementation of the current EHR-system. There was a significantly higher proportion of both former and current smokers in the LC group

compared to the non-LC group (58.8% vs. 43.0% and 34.2% vs. 26.0%, respectively, p<0.001 for all) (**Table 5**). The manual annotation was validated against the subpopulation registered in the DLCR. Among the 9,399 LC patients with available smoking status from the DLCR, the was agreement between the two registrations in 83% of non-smoker cases and 97% of current/former smoker cases.

Table 5: Smoking status

Smoking status, no.(%)	Non-LC	LC	P-value
Total	**18,287 (100)**	**2,719 (100)**	
Never smoker	5,682 (31.1)	421 (8.9)	<0.001
Former smoker	7,858 (43.0)	2,683 (58.8)	
Current smoker	4,747 (26.0)	1,615 (34.2)	

Consultations and CRP point-of-care test in general practice

Ten percent of the population did not have any recorded consultations or CRP point-of-care test within the 6-month interval before the index date. The LC group had more frequent GP visits compared to the non-LC group, although both groups had a median of 3 visits (IQR 2-6 for LC vs. IQR 2-7 for non-LC, p<0.001). Additionally, 11.2% of non-LC patients did not have any consultations compared to 8.3% of LC patients (**Table 6**). A higher percentage of LC patients had more than 4 consultations compared to the non-LC group (45.0% vs. 41.6%, p<0.001).

There was no significant difference in the distribution of CRP point-of-care tests between the two groups, with 52% in both groups not undergoing the test. However, high-stage LC patients were more prevalent in the group that underwent a CRP point-of-care test compared to those who did not (75% vs. 70%, p<0.01).

Table 6: Number of consultations and CRP point-of-care test in general practice

General practice data	Non-LC no. (%)	LC no. (%)	P-value
Total	**27,660 (100)**	**11,284 (100)**	
Consultations			
0	3,093 (11.2)	932 (8.3)	<0.001

1-4	13,053 (47.2)	5,275 (46.8)	
>4	11,514 (41.6)	5,077 (45.0)	
CRP point-of-care test			
0	14,516 (52.5)	5,973 (52.9)	
1-4	12,299 (44.5)	5,014 (44.4)	
>4	845 (3.1)	297 (2.6)	0.074

Laboratory results

Of all included patients, 34,129 (88%) had some representation from 180 days before to 14 days after the index date. When narrowing the criteria to 28 days before to 14 days after the index date and focusing solely on diagnostic LC units, the cohort was reduced to 18,462 individuals. Among these, 14,957 had at least 17 analyses performed. Although most median values were within the clinical standard reference interval, minor significant differences were found between the two groups (**Table 7**). Median values of white blood cells (leukocytes, neutrophils, monocytes), calcium, platelets, CRP, alkaline phosphatase and LDH were significantly higher in the LC group compared to the non-LC group. Similarly, median values of hemoglobin, albumin, lymphocytes, eosinophils, ALAT, creatinine, and sodium were significantly lower in the LC group compared to the non-LC group. The neutrophil-to-lymphocyte ratio was significantly higher in the LC group compared to the non-LC group (3.4 vs. 2.6, p<0.001).

Table 7: Laboratory results displayed in median values with interquartile ranges. The number of decimals varies between analyses as provided by the laboratory.

Laboratory data	**Non-LC no. (%)**	**LC no. (%)**	**P-value**
	10,503 (100)	**4,454 (100)**	
B-Hemoglobin, mmol/L	8.70 (8.0-9.3)	8.40 (7.7-9.0)	<0.001
B-Leucocytes, 10^9/L	7.64 (6.20-9.46)	9.12 (7.43-11.20)	<0.001
B-Neutrophils, 10^9/L	4.70 (3.58-6.20)	6.10 (4.71-7.95)	<0.001
B-Lymphocytes, 10^9/L	1.81 (1.39-2.33)	1.74 (1.30-2.27)	<0.001
NLR	2.6 (1.8-3.8)	3.4 (2.4-5.2)	<0.001

B-Monocytes, 10^9/L	0.65 (0.51-0.84)	0.76 (0.59-0.97)	<0.001
B-Basophils, 10^9/L	0.04 (0.02-0.06)	0.04 (0.02-0.06)	<0.001
B-Eosinophils, 10^9/L	0.17 (0.10-0.27)	0.14 (0.07-0.25)	<0.001
B-Platelets, 10^9/L	272 (223-334)	311 (250-391)	<0.001
P-Albumin, g/L	43 (41-45)	42 (39-44)	<0.001
Total Calcium, mmol/L	2.34 (2.27-2.41)	2.36 (2.29-2.43)	<0.001
P-CRP, mg/L	3.7 (1.4-10.0)	9.9 (3.0-32.0)	<0.001
P-ALAT, U/L	22 (16-31)	18 (13-26)	<0.001
P-LDH, U/L	192 (169-221)	214 (182-257)	<0.001
P-Alkaline phosphatase, U/L	75 (62-92)	83 (68-102)	<0.001
P-Bilirubin-total, µmol/L	7 (6-10)	7 (5-9)	<0.001
P-Amylase (pancreatic), U/L	25 (19-34)	25 (18-34)	0.79
P-INR	1.0 (0.9-1.1)	1.0 (0.9-1.1)	<0.001
P-Creatinine, mmol/L	76 (64-89)	72 (60-87)	<0.001
P-Sodium, mmol/L	140 (138-142)	139 (136-141)	<0.001
P-Potassium, mmol/L	4.0 (3.8-4.3)	4.0 (3.8-4.3)	0.08

Symptoms, familiar predisposition and exposure

Complete data on the aforementioned datasets were available for 9,940 individuals, of which outpatient records were accessible for 5,587. Symptoms were not reported for 10% of the LC group, in contrast to 13% of the non-LC group (p=0.002). The most frequent symptoms in both groups included cough (53.4%), dyspnea (36.3%), weight loss (25.2%), fatigue (19.9%), and haemoptysis (16.2%). The prevalence of weight loss, fatigue, back pain, and other pains was higher in the LC group, while haemoptysis and fever were more frequent in the non-LC group (p<0.001 for all) (**Table 8**). Among low-stage LC patients, 17.4% experienced none of the symptoms compared to only 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<0.001 for all) (**Figure 14**).

Familial predispositions to LC were reported in 9.0% of the LC group compared to 6.8% in the non-LC group (p=0.003). Exposure to environmental LC risk factors was present in 20.4% of all patients, with no significant difference across groups (p=0.091).

Table 8: Symptoms, familial predisposition and exposure by lung cancer (LC) state

Symptoms, familial predisposition and exposure	**Non-LC no. (%)**	**LC no. (%)**	**P-value**
Total	**3,733 (100)**	**1,854 (100)**	
Predisposition	253 (6.8)	167 (9.0)	0.00
Exposure	785 (21.0)	354 (19.1)	0.09
Hemoptysis	694 (18.6)	212 (11.4)	<0.001
Pneumonia	671 (18.0)	303 (16.3)	0.13
Cough	2,012 (53.9)	969 (52.3)	0.25
Dyspnea	1,365 (36.6)	663 (35.8)	0.56
Fever	286 (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)	0.16
Hoarseness	174 (4.7)	92 (5.0)	0.62
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)	0.01
Headache	144 (3.1)	65 (3.5)	0.37
Dizziness	161 (4.3)	96 (5.2)	0.15
Edema	196 (5.3)	108 (5.8)	0.37

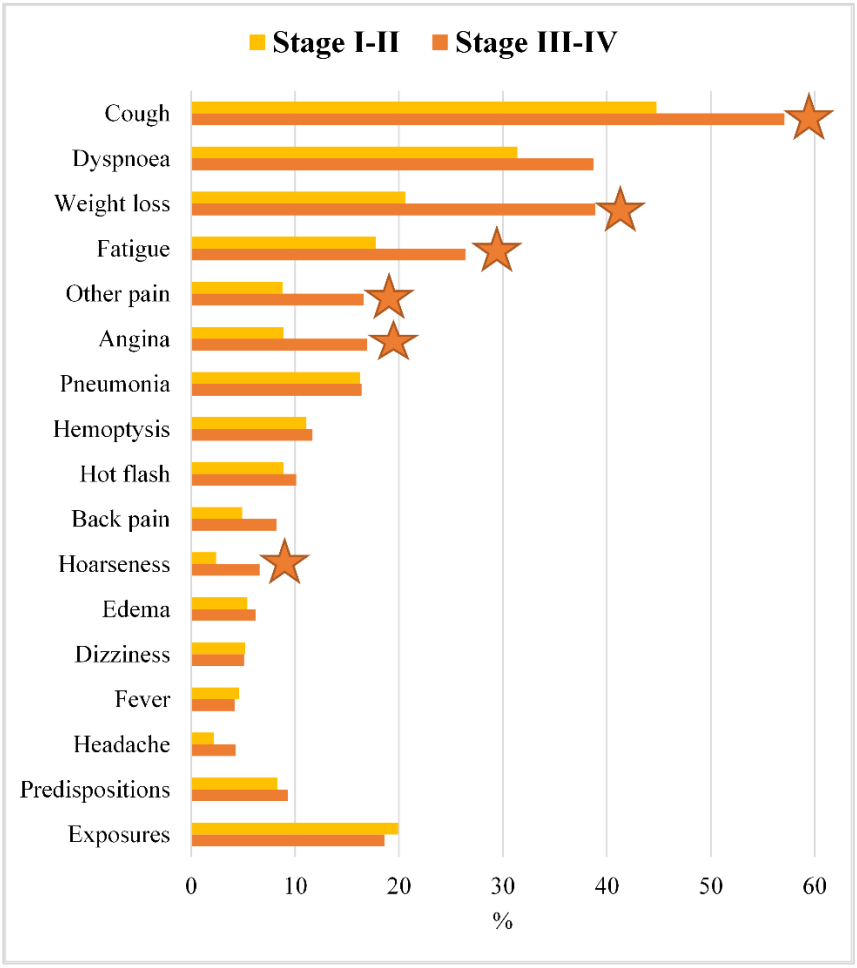


Figure 14: Symptoms, familial predisposition and exposure divided into low-stage (I-II) and high-stage (III-IV) lung cancer (LC). Significant differences are indicated with a star coloured according to the group with the highest rate.

Combined data availability

The initial cohort of 38,944 individuals underwent several reductions due to specified criteria on laboratory data (requiring a minimum of 17 analyses from diagnostic LC units within 28 days before to 14 days after the index date) and missing free text. The cohort with complete information from all data types comprised 5,587 individuals, among whom 1,854

had LC (33%) and 3,733 did not (67%). Comparison between the reduced and initial cohorts revealed an overall pattern of similarity in terms of age, sex, smoking status, consultations, CRP point-of-care tests, and laboratory results. However, the majority of the significant differences in comorbidity data and prescription medication were diminished as the cohort size decreased. The rate of individuals without comorbidity increased from 62% in the initial cohort to 72% in the reduced cohort. Please refer to Article 1, supplementary material for details.

Article II-III: Methods

AIM Articles II-III: Data from Article I (high-risk individuals) utilized to develop LC detection models. Article II employs smoking and laboratory data to construct an ML model, while Article III utilizes a BN model.

Study cohort and data variables

The study cohort utilized in Articles II and III is a subset derived from the initial cohort of 38,944 patients evaluated for suspected LC. A comprehensive description of this subset, including the integration of LC status data, is provided in Article I. The criteria for further filtration based on laboratory results are also detailed in the methods and results section of Article I. In summary, the reduced cohort consists of patients with at least 17 blood sample results within the period from 28 days before to 14 days after the index date. Only samples from the four LC diagnostic units were included, focusing on the 20 most common laboratory tests. Information on smoking status was obtained through manual annotation of free text as described in detail in the methods section of Article I.

Laboratory results and age were analyzed as continuous variables, while sex, LC status, and smoking status (categorized as never smoker versus former/active smoker) were analyzed as binary variables.

Article II: Development of ML models

In Article II, a prediction model was built using ML methods and conducted using Python (version 3.10), following the structured pipeline depicted in **Figure 16**:

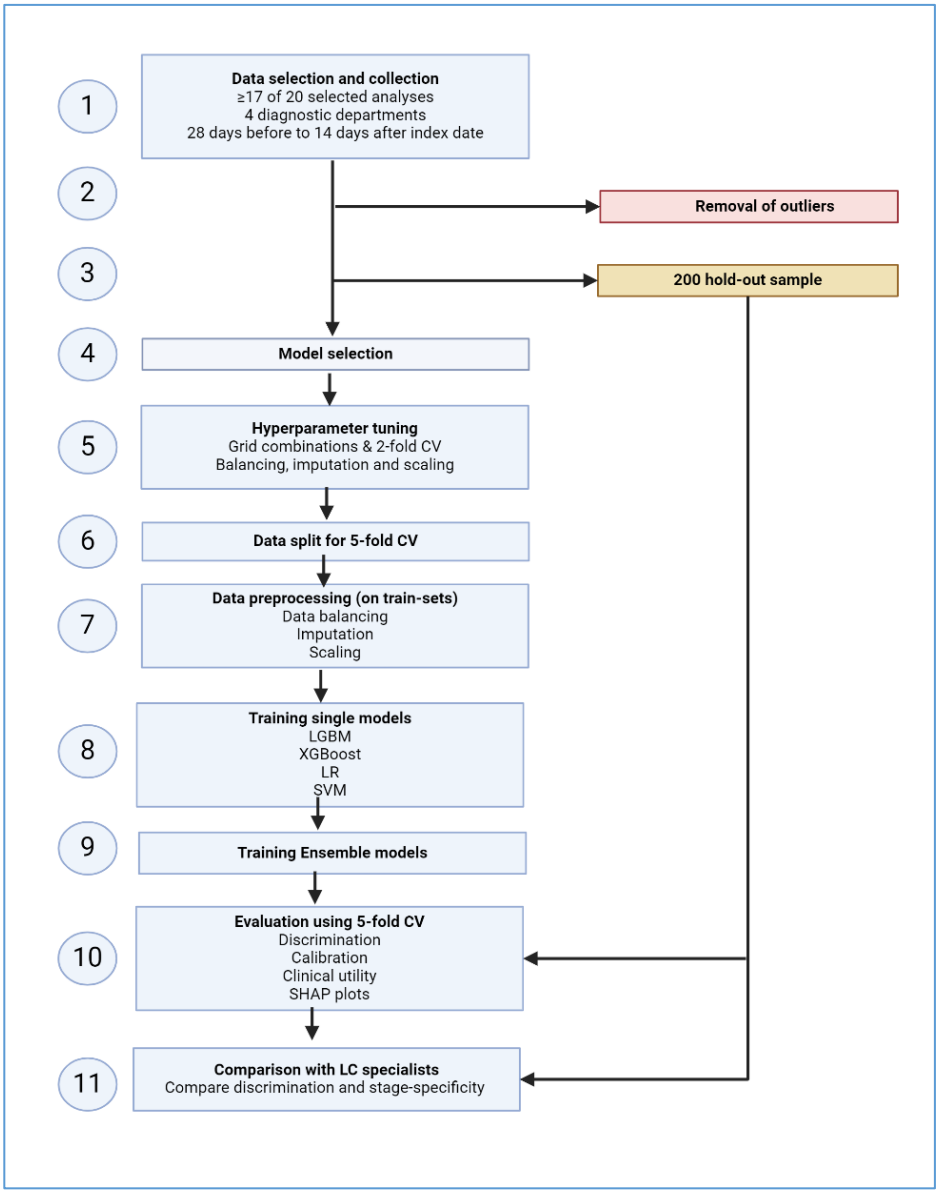


Figure 16: The ML-pipeline applied from data collection to model training and evaluation. Created with Biorender.com.

1. **Data Selection and collection:** Clinically relevant parameters known to either cause or be associated with LC were chosen. Clinical expertise guides the selection of specific analyses, diagnostic departments, and the timing for data collection. The dataset was created by applying the aforementioned filter criteria to include only individuals assessed for their risk of LC.

2. **Outlier Removal:** Extreme outliers were identified and removed from the dataset.

3. **Validation Sample:** A hold-out cohort of 200 samples sourced from Vejle Hospital in recent years and characterized by minimal missing data was set aside. It was representative of the most recent and complete data and was used for validation of the ML model and for comparison with predictions made by LC specialists given the same information.

4. **Model Selection:** We selected four well-known and relevant models: the simpler Logistic Regression (LR) and Support Vector Machine (SVM) along with the tree-based models LightGBM (LGBM) and XGBoost. The models are described in the background section.

5. **Hyperparameter Tuning:** Optimal hyperparameters were determined before model training. Relevant hyperparameters for each specific model type were identified and their ranges were specified (e.g. the number of leaves in an LGBM model ranging from 100 to 1000). A grid of all possible combinations of hyperparameters was created. For example, tuning three hyperparameters, each with five values, results in 3 x 5 = 15 possible combinations. Each combination was then used to train and evaluate the model using 2-fold cross-validation. The combination of hyperparameters that achieved the best performance was selected.

6. **Data Split:** Data was split using 5-fold cross-validation. Preprocessing steps were applied only to the training folds to avoid data leakage, ensuring that no information from the test set influenced model creation.

7. **Data Preprocessing**: This step involved data balancing, imputation, and scaling performed solely on the training folds:

- Data Balancing: To address the imbalance in the dataset, with the majority of individuals not having LC, various techniques were evaluated. The RandomUnderSampler technique, which randomly under-samples the majority class to balance the distribution between LC and non-LC patients, was chosen for its performance.

- Imputation: Three of the 20 analyses were missing, accounting for approximately 3% of the total data. Six imputation methods were evaluated, with the Friedman test showing no significant difference in F1-measure across them. Given the skewed distribution in most laboratory results and the simplicity of the median method, missing values were imputed using the median of the corresponding feature columns.

- Data Scaling: Standardization was applied to ensure proportional contribution to the model by each feature. This transforms features to have a mean of 0 and a standard deviation of 1, preventing features with larger magnitudes from dominating smaller, potentially more important features.

8. **Training single models**: The four chosen ML models were trained using the previously established optimal hyperparameters. Stratified 5-fold cross-validation was employed to maintain the same ratio of LC to non-LC patients in each training and validation fold. They were furthermore calibrated to ensure that prediction probabilities reflected observed outcomes.

9. **Training Ensemble Models:** Ensemble methods (models that combine predictions from several single models) were explored to improve robustness and performance by combining the strengths of different models. After training the four models individually, their predictions were combined to form the ensemble prediction using various dynamic ensemble selection (DES) methods. Please see Article 1, supplementary material for further details.

10. **Evaluation using 5-fold cross-validation:** Model performance was evaluated using 5-fold cross-validation in a two-step process (**Figure 17**). First, five iterations were run with a rotating single fold serving as the validation set. Second, five iterations were performed using the 200 hold-out samples instead of the validation set. This approach allowed us to report performance on both the larger dataset and the 200 hold-out set. The results of each iteration were averaged for both the validation set and the 200 hold-out samples. This evaluation was conducted for all four models individually as well as for the ensemble models.

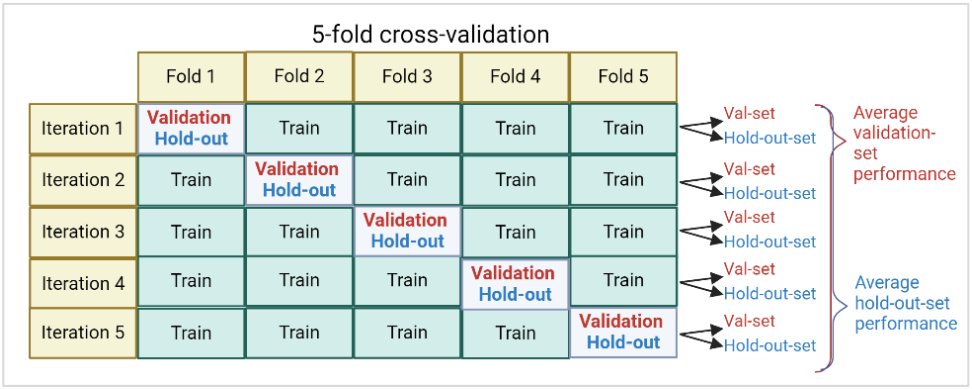


Figure 17: The 5-fold cross-validation approach with shifting validation and hold-out-sets resulting in an average validation-set performance and average hold-out-set performance.

The discriminative performance was evaluated using sensitivity, specificity, PPV, and F1-score metrics, all reported at the threshold after calibration corresponding to a pre-calibration threshold of 0.5. ROC curves were employed to compare the AUC for different models, accompanied by standard deviations (SD). Model performances were compared using the Nemenyi test with statistical significance set at a two-sided p-value of less than 0.05.

Model calibration was assessed by comparing predicted probabilities with the observed fraction of LC patients. Decision curve analyses were performed to determine the clinical net benefit as compared to default strategies of examining all or no patients. SHAP summary plots were created to illustrate the importance of different variables, and force plots were used for explainability at the individualized level. In subgroup analysis, we stratified by LC stage and created reduced models that included only the most important features as determined by the SHAP analyses.

11. **Comparison with LC specialists:**

The best-performing model from the 5-fold cross-validation was selected for comparison with predictions from the five LC specialists. Average predictions from the results of the 200 hold-out samples were used to produce a ROC curve for the model. Results from the five LC specialists (pulmonologists) were shown both as individual votes and as majority votes, each corresponding to a specific point on the ROC curve. The specific point representing the majority vote was compared to the model's performance by fixing one axis

and reporting the value on the other axis. Finally, stage-specific performance was compared with predictions made by the specialists.

Article III: Development of BN models

In Article III, prediction models were constructed using BNs. The experimental setup included three different types of analyses, resulting in 16 distinct models: varying degrees of missing data, expert-drawn DAGs versus data-learned DAGs, and standard versus auto-generated discretization of laboratory variables. All models were validated using 10-fold cross-validation with 95% confidence intervals for the AUC. Discrimination, calibration, and clinical utility were assessed as in Article II. Although SHAP plots are not compatible with Bayesian Networks, the DAGs were provided to facilitate explainability. All experiments were conducted using the WEKA analytical framework (version 3.8) depicted in **Figure 18** and described below.

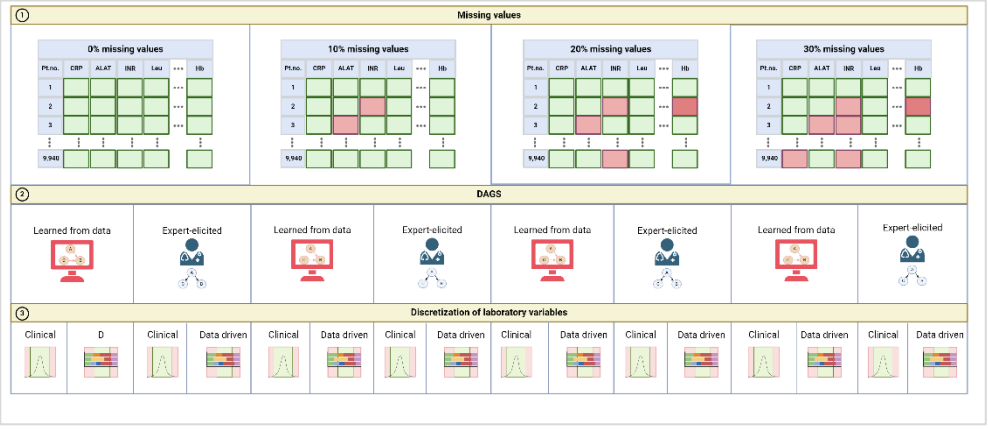


Figure 18: The experimental setup involving different degrees of missing data, expert-elicited DAGS versus DAGs learned from data, and clinical versus data-driven discretization strategies.

1. **Different degrees of missing values:** Missing data were artificially introduced into the dataset by randomly deleting values. The level of missing values was set to 0%, 10%, 20% and 30%.

2. **DAGs Learned from Data versus Expert-Elicited DAGs:** BNs were trained using two different approaches for constructing DAGs. The first one used an “expert-elicited

DAG", which was created with input from three clinicians specializing in oncology, pulmonology, and biochemistry, respectively.

The second approach used a "DAG learned from data", constructed using the K2 algorithm [115]. This is a fundamental method used in probabilistic graphical models like BNs. Named after the K2 mountain, it compares the challenges of learning BNs from data with climbing a tall mountain. The K2 algorithm was used for structure-learning to find the optimal BN structure. It assumed complete data, replacing missing continuous attributes with the mean, and missing discrete attributes with the mode. The structure of the K2-DAG was built by adding one connection at a time, assessing each added connection's predictive performance in establishing the most representative DAG of probabilistic relationships among variables. The K2 algorithm was tested with varying limits on parent nodes per variable in the DAG, from 1 to 10. The best structure was determined based on its AUC and validated through 10-fold cross-validation.
In the parameter learning phase of both the expert-elicited DAG and the DAG learned from data, specific probabilities established between the variables were determined using the expectation maximization algorithm. This technique handles missing data by estimating missing values in the expectation step and refining the model in the maximization step, ultimately enhancing model performance [116].

3. **Clinical versus data driven discretization of laboratory variables:** In BNs, continuous variables are often converted into discrete (categorical) variables. This allows for the use of established inference and learning methods designed for discrete variables [117]. In clinical practice, laboratory results are typically evaluated using the standard reference interval based on the 95% confidence interval. This generally categorizes results into three bins: below reference, within reference, and above reference. We compared clinical discretization, which uses standard reference intervals, with a data-driven approach based on the minimum description length strategy [118]. The strategy seeks to identify the model that minimizes the information needed to describe the data. It involves selecting the optimal number of bins for continuous laboratory variables, striking the best trade-off between model simplicity and accuracy in representing the underlying data distribution.

Article II-III: Results

Descriptive characteristics

After applying the filter criteria to the initial cohort of 38,944 patients suspected of having LC, 9,940 individuals had at least 17 of 20 analyses present, representing the four diagnostic units within the period of 28 days before to 14 days after the index date, and available smoking information. **Table 9** displays the clinical and laboratory results of this cohort, i.e. 2,505 (25%) LC patients and 7,435 non-LC patients. The median age of LC patients was 75 years (IQR 68-80) compared to 71 years in non-LC individuals (IQR 59-79). The LC and non-LC group comprised 52% and 44% females, respectively. Among LC patients, 92% were current or former smokers compared to 69% of the non-LC patients.

The LC group had significantly higher median values for white blood cells (leukocytes, neutrophils, monocytes), calcium, platelets, CRP, alkaline phosphatase, and LDH compared to the non-LC group. Similarly, the median values for hemoglobin, albumin, lymphocytes, eosinophils, ALAT, creatinine, and sodium were significantly lower in the LC group compared to the non-LC group. Clinical demographics and median values for blood test results are detailed in Table 11. This dataset was used for development of prediction models in both Article II and III.

Table 9: Clinical demographics and laboratory results of the 9,940 included individuals. Data are reported in counts and percentages or medians and interquartile ranges. U/L: units per liter; g/L: milligrams per liter; 109/L: count of cell type 109/L per liter; mmol/L: millimoles per liter; mmol/L: micromoles per liter. P-: Plasma. B-: Blood. ALAT: alanine aminotransferase; CRP: c-reactive protein; INR: international normalized ratio; LDH: lactate dehydrogenase. The number of digits reported on the laboratory results reflects the number of digits provided by the laboratory.

Variable	Reference interval	Non-LC (n=7,435)	LC (n=2,505)
Age, years		71 (59-79)	74 (68-80)
Sex			
Female		3,273 (44.0%)	1,304 (52.1%)
Male		4,162 (56.0%)	1,201 (47.9%)
Smoking status			
Never smoker		2,288 (30.8%)	196 (7.8%)

Former/current smoker		5,147 (69.2)	2,309 (92.2%)
Blood sample analyses			
ALAT, U/L	Male: 10-70, Female: 10-45	22 (15)	19 (12)
Albumin, g/L	34-45	43 (4)	42 (5)
Alkaline phosphatase, U/L	10-65	25 (15)	25 (15)
Alkaline phosphatase	35-105	74 (29)	81 (32)
Basophils, 10^9/L	<0.02	0.04 (0.04)	0.05 (0.04)
Bilirubin-total, µmol/L	5-25	7 (5)	7 (4)
CRP, mg/L	<6	3.4 (7.9)	7 (19.7)
Calcium-total, mmol/L	2.15-2.51	2.34 (0.13)	2.38 (0.14)
Eosinophils, 10^9/L	<0.05	0.17 (0.27)	0.14 (0.16)
Hemoglobin, mmol/L	Male: 8.3-10.5, Female: 7.3-9.5	8.7 (1.2)	8.5 (1.3)
INR	<1.2	1 (0.15)	1 (0.14)
Potassium, mmol/L	3.5-4.4	4 (0.5)	4 (0.5)
Creatinine, mmol/L	Male: 60-105, Female: 45-90	76 (26)	72 (26)
LDH, U/L	115-255	192 (52)	209 (64)
Leucocytes, 10^9/L	3.5-8.8	7.62 (3.18)	8.8 (3.41)
Lymphocytes, 10^9/L	1.0-4.0	1.84 (0.97)	1.79 (0.97)
Monocytes, 10^9/L	0.2-0.8	0.65 (0.32)	0.73 (0.36)
Sodium, mmol/L	137-145	140 (4)	139 (5)

Neutrophils, 10^9/L	1.5-7.5	4.66 (2.57)	5.77 (2.90)
Platelets, 10^9/L	Male: 145-350, Female: 165-390	271 (107)	301 (135)

Article II: Evaluation of ML-models

Discrimination, calibration and clinical utility

The performance measures of the averaged validation set, derived using 5-fold cross-validation, are presented in **Table 10**. The results include the four individual models as well as the best performing ensemble model (DES). The Nemenyi post-hoc test did not indicate any statistically significant difference between the models across all performance measures. Consequently, we selected the ensemble model (DES) for further evaluation, as it robustly combines the four individual models. The DES model achieved a mean ROC-AUC of 0.77±0.01. At a risk-threshold of 0.26, corresponding to a pre-calibration threshold of 0.5, the DES model correctly classified 76.3% of the LC patients and 63.8% of the non-LC patients. At a fixed specificity of 95% the DES model demonstrated a sensitivity of 21%.

Table 10: Performance measures of the four single models and the ensemble model (DES) at a risk-threshold after calibration corresponding to a pre-calibration threshold of 0.5. Numbers are in mean values followed by standard deviations.

Model	Sensitivity	Specificity	Positive predictive value	F1-score	ROC-AUC
LGBM	76.2 ± 2.6	63.4 ± 2.7	41.3 ± 1.2	53.6 ± 0.8	76.9.0 ± 0.6
XGBoost	75.3 ± 1.9	64.2 ± 3.0	41.6 ± 1.7	53.6 ± 1.3	76.5 ± 1.0
LR	72.5 ± 3.5	66.1 ± 4.0	42.1 ± 1.8	53.2 ± 0.7	75.8 ± 0.7
SVM	74.5 ± 3.4	63.7 ± 3.9	41.1 ± 1.5	52.9 ± 0.5	75.7 ± 0.8
DES	76.2 ± 2.0	63.8 ± 2.2	41.5 ± 1.1	53.8 ± 0.9	76.8 ± 0.9

Calibration was assessed using the calibration curve shown in Figure **19A**. Prior to calibration, the model systematically overestimated the risk of LC. After calibration,

however, it showed good alignment, with predicted probabilities closely matching the actual proportion of LC patients, though. The Brier score was 0.15, indicating a moderately well-calibrated model, though there remains room for improvement, likely due to slight over- and underestimations in the higher prediction intervals.

Decision curve analysis is illustrated in Figure **19B**. The graph provides insight into the clinical utility at different risk thresholds. It shows the net benefit of using the DES model for classifying high-risk LC patients compared to the strategies of selecting all patients (grey line) or no patients (blue line). The DES model demonstrates a higher net benefit across threshold probabilities from approximately 7% to 70% compared to the other two clinical strategies.

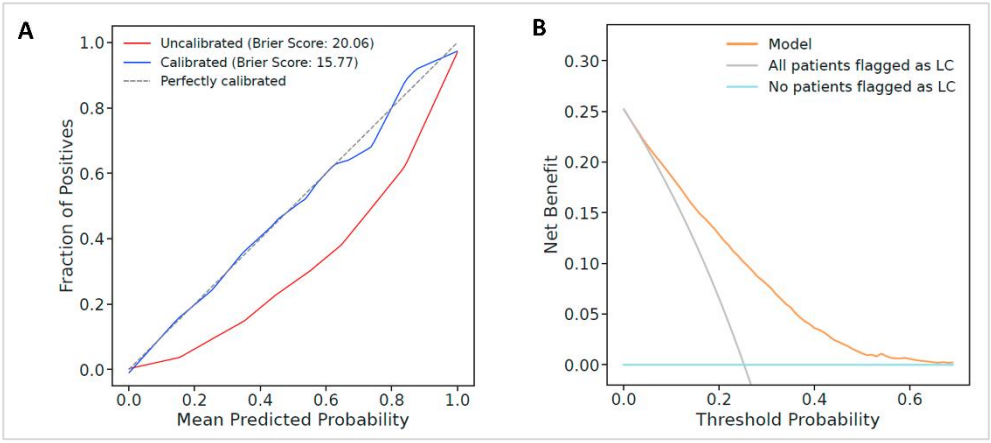


Figure 19: Calibration assessed through calibration curve (A). The predicted probability is shown on the x-axis, the observed LC fraction on the right y-axis. Clinical utility plot (B) displaying the net benefit when using the DES-model for LC prediction compared to the two extreme scenarios of selecting all patients to be at high risk (grey line) or no patients (blue line). Both graphs display the results of the DES-model.

Explainability

The SHAP plots as well as the feature removal plot of the DES-model is presented in **Figure 20**. **Figure 20A** shows the summary plot with the 10 most important features of the DES model ranked by importance. The most crucial feature is a "high" smoking value (active or former smoker) followed by elevated levels of LDH, high age, high calcium and low sodium.

Figure 20B shows the F1-performance as the number of features included in the model is reduced, based on the SHAP force-plot analysis of feature importance. The DES model achieved its highest F1-measure with the top 10 features, with no improvement when including more than 10. Model performance decreased by only 2% when reduced to the top 5 features.

Figures 20C and **D** display SHAP force plots illustrating an individual's risk of LC based on their specific values. In **Figure 20C,** the individual is a true positive case (LC patient correctly predicted as LC) with ALAT and total-calcium being the most significant contributors to the predicted high risk. **Figure 20D** represents a false positive case (non-LC patient incorrectly predicted as LC), where low sodium and ALAT levels, together with high age were the main features driving the high LC prediction.

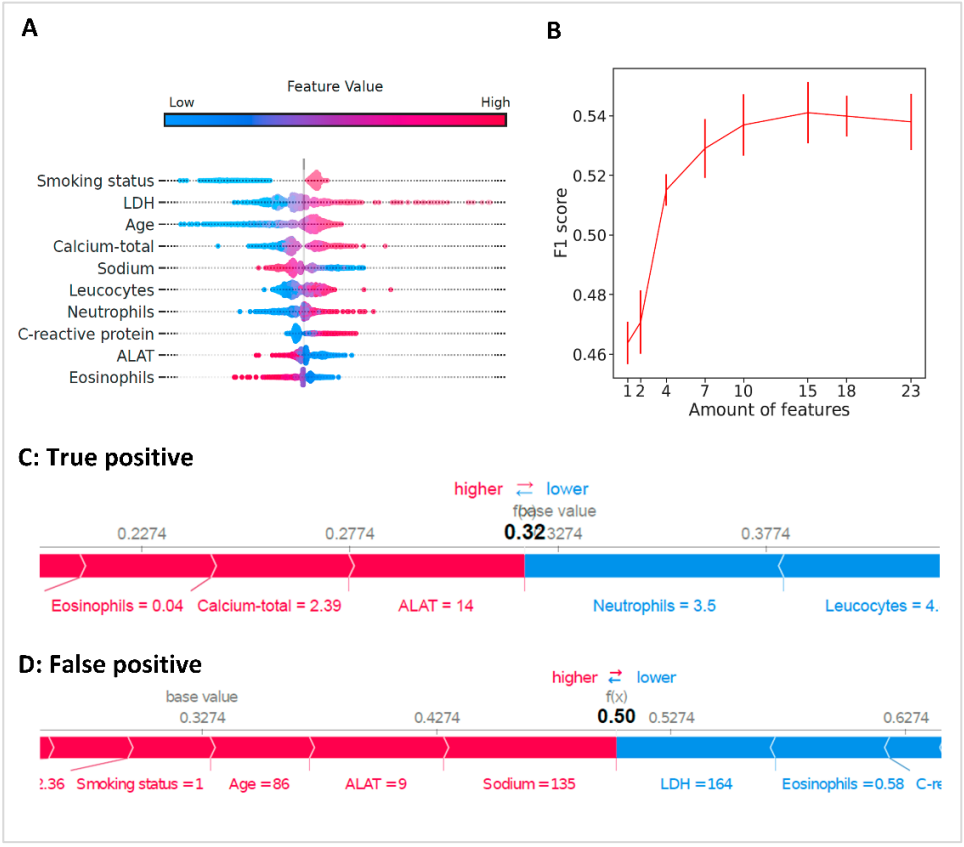


Figure 20: SHAP summary plot (**A**) depicting only the top-10 features which are most influential for the DES-model listed after order of importance. Feature removal plot (**B**) comparing the F1-measure

(y-axis) when reducing the number of features included in the model. The horizontal lines represent SD from the 5-fold cross-validation. SHAP force plots (**C, D**) displaying the risk factors contributing to the exact prediction on an individual level. All plots display the results of the DES-model.

Comparison with LC specialists on 200 hold-out sample

The performance of the DES model is compared with that of LC specialists on the ROC curve shown in **Figure 21A.** The DES model, evaluated using 5-fold cross-validation, achieved an AUC of 0.80±0.01. The specialist performance based on a majority vote for each of the 200 samples resulted in a sensitivity of 67.4% and a specificity of 70.3%. At the same level of specificity, the DES model demonstrated superior sensitivity achieving 73.8% on the same 200 patients. **Figure 21B** illustrates the proportion of LC patients at each stage for both the average specialist and the DES model compared to the actual proportions. The DES model outperformed the specialists in diagnosing stage I and III LC patients, while their performance was similar for stage II. However, the specialists were better at diagnosing stage IV LC patients. The ability to detect LC patients in stage I is particularly important since symptoms and large patterns in laboratory analyses often only arise later. This highlights the potential of AI methods to assist clinicians in detecting subtle differences that may fall within reference intervals and are therefore not easily noticeable by clinicians.

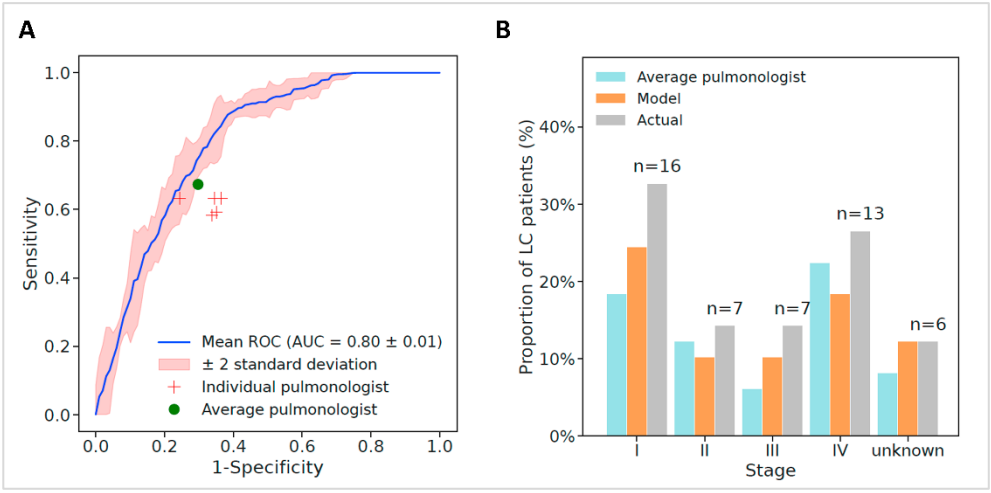


Figure 21: ROC curve representing average performance of the DES model assessed using 5-fold cross-validation with a hold-out sample of 200 (**A**). Blue line=DES model, red crosses=individual predictions from the five LC specialists, green mark= average measure of the specialists. **B:** Stage-specific proportion of LC patients by DES model, average of specialists, and the actual proportion.

Article III: Evaluation of BN models

Discrimination, calibration and clinical utility

The combination of missing values, DAG structures, and discretization strategies resulted in the development of 16 models. **Table 11** presents specific evaluation metrics for these models, with sensitivity and specificity reported at a default risk threshold of 0.5. Overall, the AUCs remained consistent regardless of the level of missing values, ranging from 0.74 (95% CI: 0.71-0.76) for Model 11 to 0.76 (95% CI: 0.73-0.78) for Model 6. There was no significant difference between expert-elicited DAGs and DAGs learned from data; nor was there a noticeable difference in AUCs between standard and data-driven discretization strategies.

At a fixed TNR of 95%, the highest TPR was achieved by Model 2 at 20.6% (95% CI: 19.9%-21.3%). It also had an AUC of 0.76 (95% CI: 0.73-0.78), which is slightly lower than the DES model from article II, which had an AUC of 0.77 and a TPR of 21% at a fixed TNR of 95%.

Table 11: Evaluation metrics of the 16 models developed from the combination of missing values (0-30%), DAG structure (learned from data versus expert-elicited) and discretization strategy (clinical versus data driven).

Missing values	DAG	Discreti-zation	Model no.	Sensitivity (95%CI)	Specificity (95% CI)	AUC (95% CI)
0%	Learned from data	Clinical	**1**	0.37 (0.36-0.37)	0.89 (0.87-0.91)	0.76 (0.73-0.78)
		Data driven	**2**	0.34 (0.34-0.35)	0.90 (0.88-0.91)	0.76 (0.73-0.78)
	Expert-elicited	Clinical	**3**	0.44 (0.44-0.45)	0.83 (0.82-0.85)	0.74 (0.71-0.76)
		Data driven	**4**	0.37 (0.36-0.77)	0.88 (0.86-0.89)	0.75 (0.72-0.77)
10%	Learned from data	Clinical	**5**	0.33 (0.32-0.33)	0.90 (0.88-0.91)	0.75 (0.72-0.78)
		Data driven	**6**	0.31 (0.31-0.32)	0.91 (0.89-0.92)	0.76 (0.73-0.78)

	Expert-elicited	Clinical	**7**	0.42 (0.41-0.43)	0.85 (0.83-0.86)	0.74 (0.71-0.77)
		Data driven	**8**	0.34 (0.33-0.35)	0.88 (0.87-0.90)	0.75 (0.72-0.77)
20%	Learned from data	Clinical	**9**	0.30 (0.30-0.31)	0.90 (0.88-0.92)	0.75 (0.72-0.78)
		Data driven	**10**	0.27 (0.27-0.28)	0.91 (0.89-0.92)	0.75 (0.72-0.77)
	Expert-elicited	Clinical	**11**	0.37 (0.37-0.38)	0.85 (0.84-0.87)	0.74 (0.71-0.76)
		Data driven	**12**	0.31 (0.30-0.31)	0.89 (0.87-0.91)	0.74 (0.72-0.77)
30%	Learned from data	Clinical	**13**	0.25 (0.24-0.25)	0.92 (0.91-0.94)	0.75 (0.73-0.78)
		Data driven	**14**	0.27 (0.27-0.28)	0.91 (0.89-0.93)	0.75 (0.72-0.77)
	Expert-elicited	Clinical	**15**	0.34 (0.33-0.35)	0.87 (0.86-0.89)	0.74 (0.71-0.77)
		Data driven	**16**	0.27 (0.27-0.28)	0.90 (0.89-0.92)	0.75 (0.72-0.77)

Model calibration was assessed using the predicted versus observed plot shown in **Figure 22A**. It illustrates four models all employing DAGS learned from data with data-driven discretization, but each with different degrees of missing data, i.e. models 2, 6, 10, and 14. The x-axis represents the predicted risk in bins of 0.1 and the y-axis the true, observed risk. The plot demonstrates a clear correspondence between predicted and observed risks, with an increase in predicted risk followed by an increase in observed risk. For risk intervals between 0% and 40% the model is well-calibrated with predicted and observed risks falling within the same intervals. However, for risk intervals above 40% the predicted risk generally exceeds the observed risk, indicating that the model overestimates the true risk.

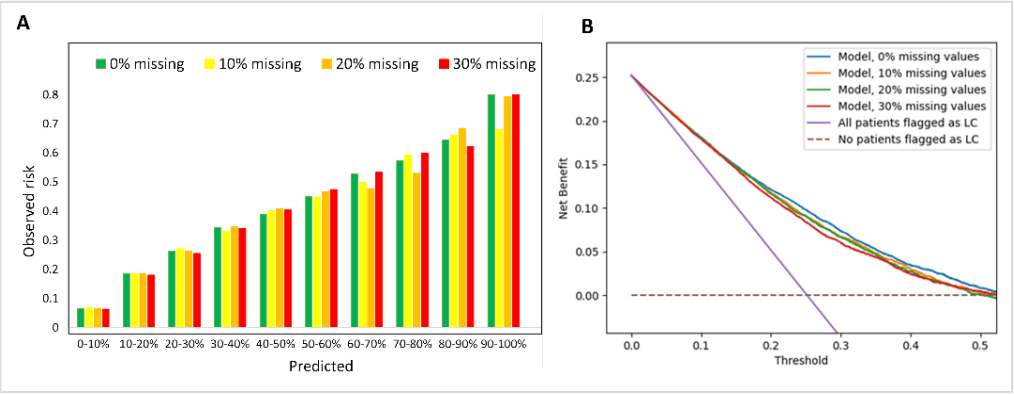


Figure 22: Figure 22A: Calibration assessed through the predicted versus observed curve. Different levels of missing data are displayed for each 10% risk interval. Figure 22B: Clinical utility plot with net benefit on the y-axis plotted against the risk threshold. The four models, each with varying degrees of missing data, are compared against the strategy of screening all patients (purple line) and no patients (brown dashed line). Both figures are based on DAGs learned from data with data-driven discretization.

The clinical utility of models 2, 6, 10, and 14 was assessed using decision curve analysis as shown in **Figure 22B**. All four models exhibit similar net benefits at lower risk thresholds and outperform the strategy of flagging all patients (purple line) when the risk threshold exceeds approximately 5%. This positive net benefit gradually declines until a risk threshold of around 50% is reached, at which point it levels off with the strategy of flagging no patients (brown dashed line). Therefore, the BN models can be considered effective for including patients in a screening scenario at a minimum threshold of approximately 5% in this population.

Explainability

The expert-elicited DAG depicted in **Figure 23** was compared to the eight DAGs learned from data, which were derived from varying degrees of missing values and different clinical versus data-driven discretization strategies (Article III, Supplementary Appendix, Figure 4). Due to the complexity of comparing DAGs with a high number of variables, a summary table was created to highlight all links (Article III, Table 4). The comparison revealed three main trends:

1. There was a general consensus between the expert-elicited DAG and the DAGs learned from data regarding links directly from LC to other variables.

2. Several links appeared only in the expert-elicited DAG, primarily connections between various laboratory variables and factors such as sex, smoking, and age.

3. The DAGs learned from data exhibited several links not present in the expert-elicited DAG. Some of them may reflect collinearity between variables such as those between leucocytes and monocytes or leucocytes and neutrophils, both of which appeared in all eight DAGs learned from data. Other notable links include those between CRP and leucocytes, and CRP and LDH.

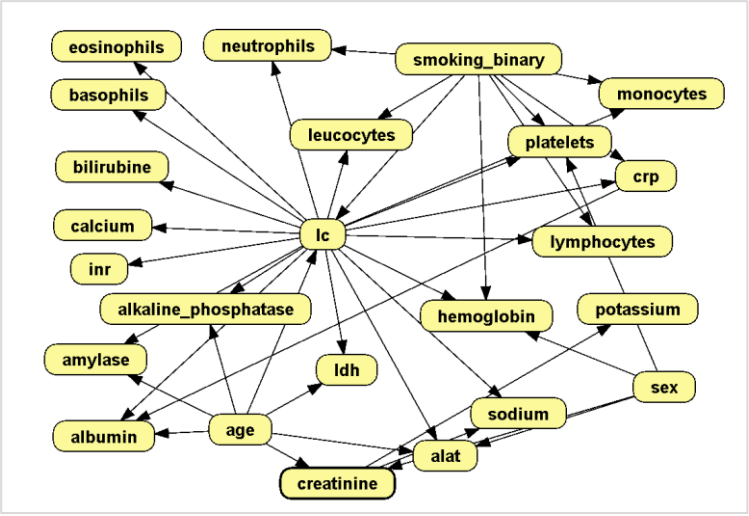


Figure 23: The directed acyclic graph (DAG) provided by experts illustrates the interconnections between the outcome, lung cancer (LC), and all other variables. This graph was created using OPEN Markov [119].

Article IV Methods

AIM Articles IV: Article IV extends the BN model to incorporate additional data types, aiming to identify the optimal model performance across data completeness and size, and the best performing combination of variables.

Study cohort and data variables

The study cohort used in Article IV comprises the entire initial cohort of 38,944 patients evaluated for suspected LC as described in Article I [1]. **Figure 24** provides an overview of the study cohort, inclusion criteria, and data sources. The variables were categorized into four datasets based on common data availability: comorbidity, laboratory results, smoking, and symptoms datasets. For detailed definitions of the data variables, please refer to Article I.

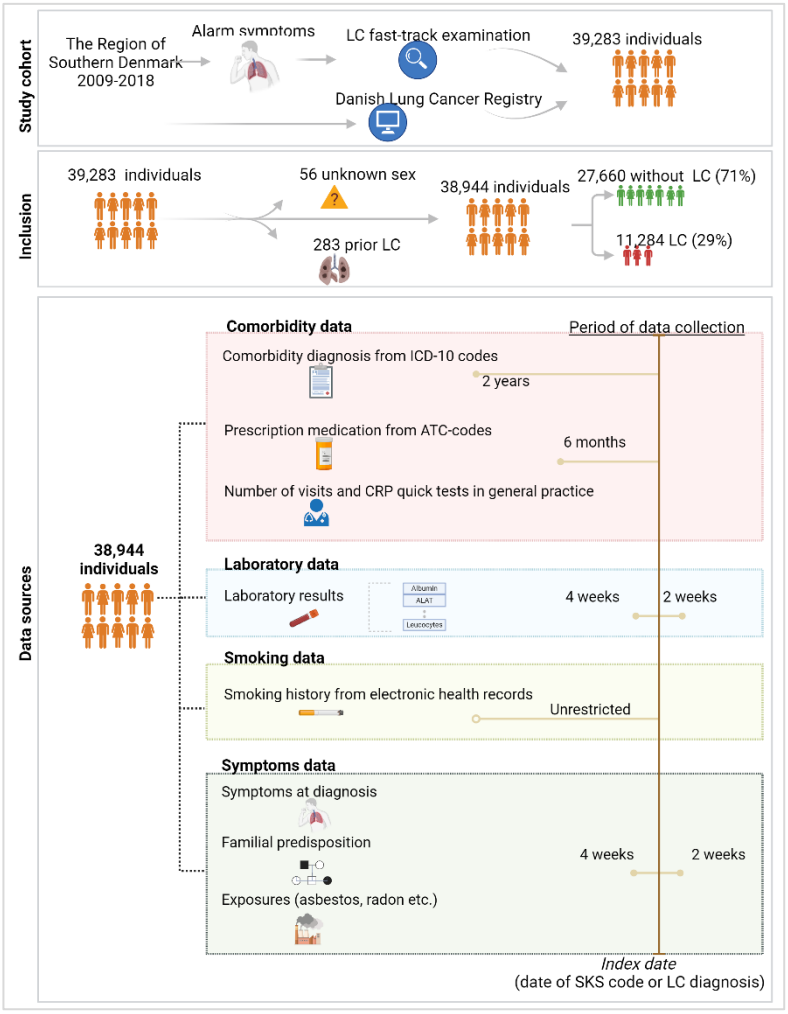


Figure 24: Study Cohort, inclusion criteria and data collection. The data were collected for specific periods leading up to the date of inclusion, referred to as the index date, and are depicted by the bars on the right side of the image. Created with Biorender.com.

Experimental setup

Before constructing the BN models, continuous variables were discretized using the MDL strategy described in Article III. The BN models were then developed using the K2 algorithm for structure learning and the expectation maximization algorithm for parameter learning. For details on these two algorithms, please refer to article II.

During the structure learning phase, missing continuous attributes were replaced with their means, while missing discrete attributes were replaced with their modes. The K2 algorithm

was tested with 1-10 number of parent nodes per variable with the best structure selected based on its AUC and validated through 10-fold cross-validation.

Some initial exploratory experiments were also conducted comparing K2 with expert drawn graphs with no interesting differences found. For the sake of simplicity, we kept K2 for the final experiments.

To investigate model performance with varying dataset sizes, completeness levels and attributes, the study cohort was divided into subsets. **Figure 25** displays Dataset A-D derived from the four data categories of comorbidity, laboratory results, smoking and symptoms. Dataset A includes individuals holding only comorbidity data, Dataset B individuals with comorbidity and laboratory results, Dataset C individuals with comorbidity, lab and smoking data, and Dataset D individuals with all four categories.

Datasets A-D were combined to reflect a real-world distribution of missing data within datasets, where certain groups of individuals have complete data and others lack information in some categories such as smoking habit or symptoms. While Dataset D represents a small but nearly complete dataset with only 2% missing data, the other dataset combinations have higher rates of missing data: Datasets CD, BCD, and ABCD have 16%, 21%, and 39% missing data, respectively. General variables such as sex and age were included in all developed models.

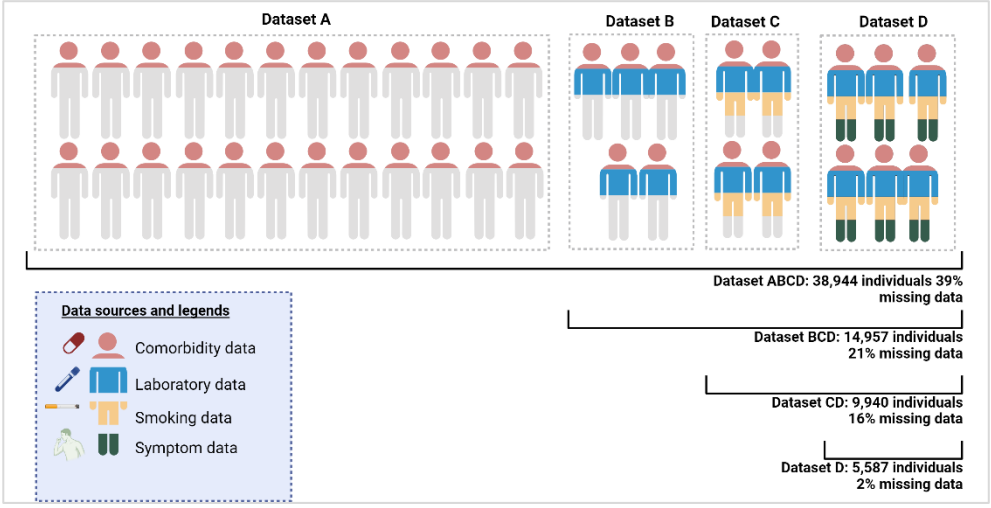


Figure 25: Datasets 1-4 and the combination of them with increasing degree of missing data. The color code of the individuals reflects the type of data available from the four combinations of data; comorbidity, laboratory results, smoking and symptoms. Created with Biorender.com.

Evaluation setup

Models were trained on the records of the four combination of datasets (ABCD/BCD/CD/D) as described in **Figure 25**. Additionally, each of the four combinations was trained using all 15 combinations of the four data categories **(Figure 26)**. One example is a model trained on dataset ABCD using only the comorbidity related variables. Validation involved a combination of 10-fold cross-validation on overlapping datasets and external validation on non-overlapping datasets. For example, a model trained on dataset D was evaluated using cross-validation within dataset D. Similarly, if the model was validated on dataset A, this served as an external validation. When larger datasets were used for validation, the results from both validation methods were integrated. For instance, the performance of a model trained on dataset D and validated on dataset ABCD was assessed by combining the cross-validation results from dataset D with the external validation results from datasets A, B, and C.

To determine the optimal set of variables, the performance of all possible combinations of the four data categories was compared using the most complete dataset (dataset D) for both training and validation through 10-fold cross-validation. **Figure 26** displays the 15 possible combinations. By comparing with the same dataset, the cohort size remained consistent , ensuring that any performance improvements were solely due to the combination of variables rather than an increase in cohort size.

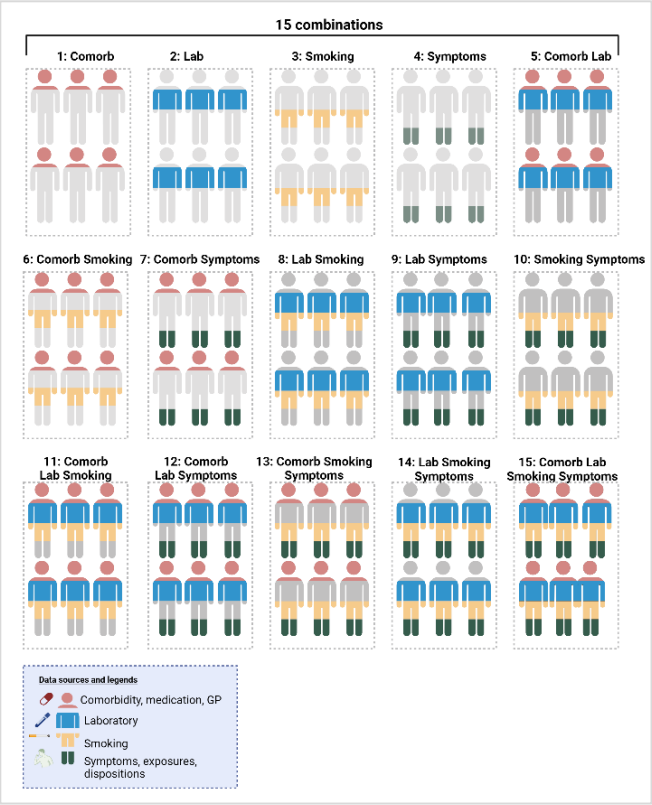


Figure 26: The 15 models trained and validated on dataset 4, comparing different combinations of the four data types: comorbidity, laboratory results, smoking history, and symptoms. Created with Biorender.com.

Article IV: Results

Table 12 presents the AUC values for models trained on datasets ABCD, BCD, CD, and D with validation conducted on datasets A, B, C, D, BCD, CD, and ABCD. For each combination, the 15 possible combinations of the four data types were evaluated, but only the top-performing combination of variables is mentioned. For the full table of results, and remaining combinations, please refer to article IV Table 2 and the supplementary material of article IV.

The best-performing model was trained on the relatively large dataset BCD, validated on dataset B, and incorporated comorbidity, laboratory results, and smoking data achieving an AUC of 0.79. This indicates that the optimal model was not derived from the largest cohort

(dataset ABCD) or the most complete dataset (dataset D), but rather from a combination of the two (dataset BCD), which included 14,957 individuals and had 21% missing data.

The second-best model was trained on the small yet nearly complete dataset D, validated on dataset D, and incorporated all four types of data variables achieving an AUC of 0.78. When validated on larger datasets with higher rates of missing data, its performance remained almost identical, with AUC values slightly decreasing to 0.77 on dataset CD (16% missing data) and 0.78 on dataset BCD (21% missing data). This shows that a model trained on high-quality data can maintain strong performance even when applied to datasets with higher rates of missing data. For instance, a model built from data on comorbidity, laboratory results, smoking, and symptoms on a small cohort (dataset 4) remains stable in performance when applied to larger cohorts with partially missing data on smoking and symptoms (dataset BCD).

This approach, however, has limitations. When validated on dataset ABCD, which contained 39% missing data, the model's performance significantly declined with the AUC dropping to 0.67. Additionally, all models validated on dataset ABCD, and particularly on dataset A, demonstrated poor performance regardless of the training set used. This indicates that there is a threshold for the amount of missing data manageable by the model, and that comorbidity data alone may not be sufficient for effective LC detection.

Table 12: AUC measures obtained from various combinations of datasets A-D used for training and validation. For each dataset, the table includes the results of testing all 15 possible combinations of data types also indicating the best combination for each case. The top two performing models are highlighted in bold. AUCs are supported by 95% confidence intervals.

	Training data ABCD	**Training data BCD**	**Training data CD**	**Training data D**
Val. data	**AUC (95%CI)**	**AUC (95%CI)**	**AUC (95%CI)**	**AUC (95%CI)**
A	0.63 (0.62-0.64)	0.62 (0.60-0.63)	0.60 (0.59-0.61)	0.60 (0.59-0.61)
B	0.78 (0.75-0.82)	**0.79 (0.75-0.83)**	0.78 (0.74-0.81)	0.77 (0.73-0.80)
C	0.72 (0.67-0.76)	0.72 (0.78-0.77)	0.73 (0.69-0.78)	0.73 (0.58-0.80)

D	0.75 (0.72-0.79	0.77 (0.73-0.80)	0.77 (0.73-0.80)	**0.78 (0.75-0.82)**
CD	0.76 (0.73-0.79)	0.77 (0.74-0.80)	0.78 (0.75-0.81)	0.77 (0.74-0.80)
BCD	0.77 (0.75-0.79)	0.78 (0.76-0.80)	0.78 (0.75-0.80)	0.78 (0.75-0.80)
ABCD	0.69 (0.68-0.70)	0.68 (0.67-0.69)	0.66 (0.65-0.67)	0.67 (0.66-0.68)

The comparison of variable combinations, assessed by training and validating on dataset D, is shown in **Figure 27**. The best of the 15 combinations was the model that included all four types of data, achieving an AUC of 0.78 (previously mentioned as the second-best model overall). In comparison, a model based on laboratory data and smoking had an AUC of 0.76, which is still superior to a model using only smoking status (along with age and gender) with an AUC of 0.70. Laboratory data and smoking were consistently included in the top-performing models, while comorbidities and symptoms appeared to have the least overall impact on model performance.

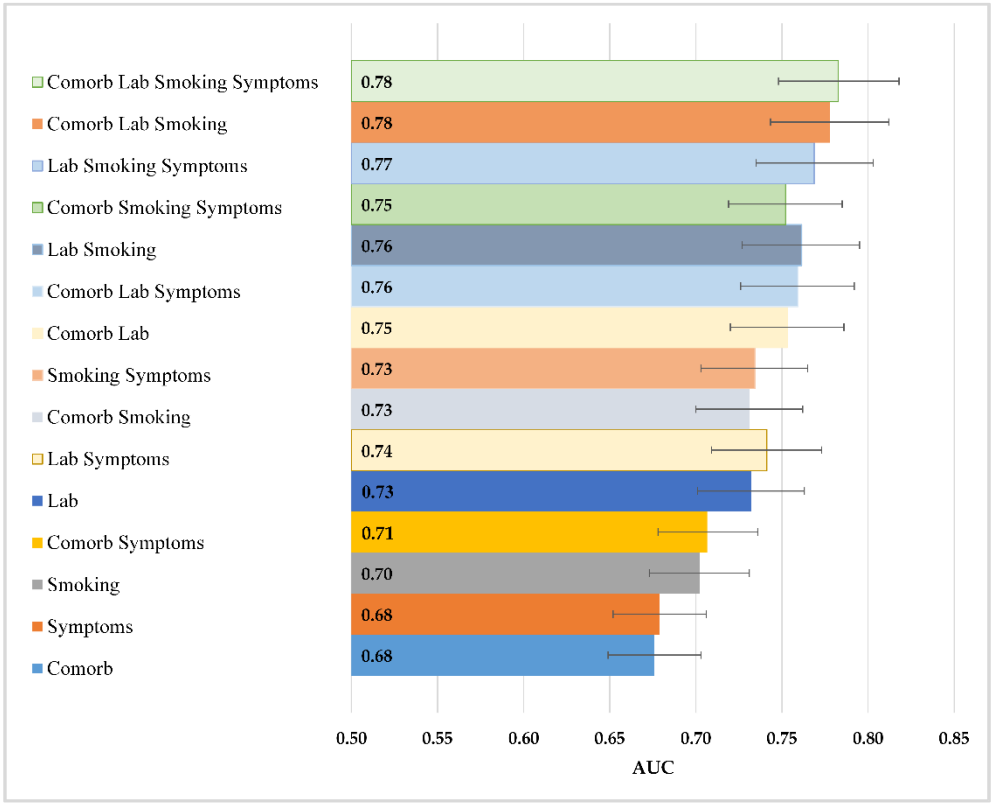


Figure 27: Comparison of AUCs when combining the four types of data in models trained and validated on dataset D (5,587 individuals).

Article V: Methods

Article V aims to assess LC incidence and stage distribution among COPD outpatients, thereby evaluating the relevance of extending LC screening models to this demographic.

Study cohort and data sources

While Article I focused on a cohort of high-risk individuals, Article V aims to investigate LC prevalence and stage distribution among a medium-risk population, specifically COPD outpatients. We conducted a retrospective analysis using data from individuals treated in the Region of Southern Denmark over a seven-year period from 1 January 2012 to 31 December 2018. Three cohorts were defined based on data from the Danish Registry of Chronic Obstructive Pulmonary Disease (DrCOPD), the Danish Lung Cancer Registry, and the regional data warehouse, which includes information on individuals examined on suspicion of LC. **Figure 28** provides a comprehensive overview of the study population along with key statistics defining the cohort.

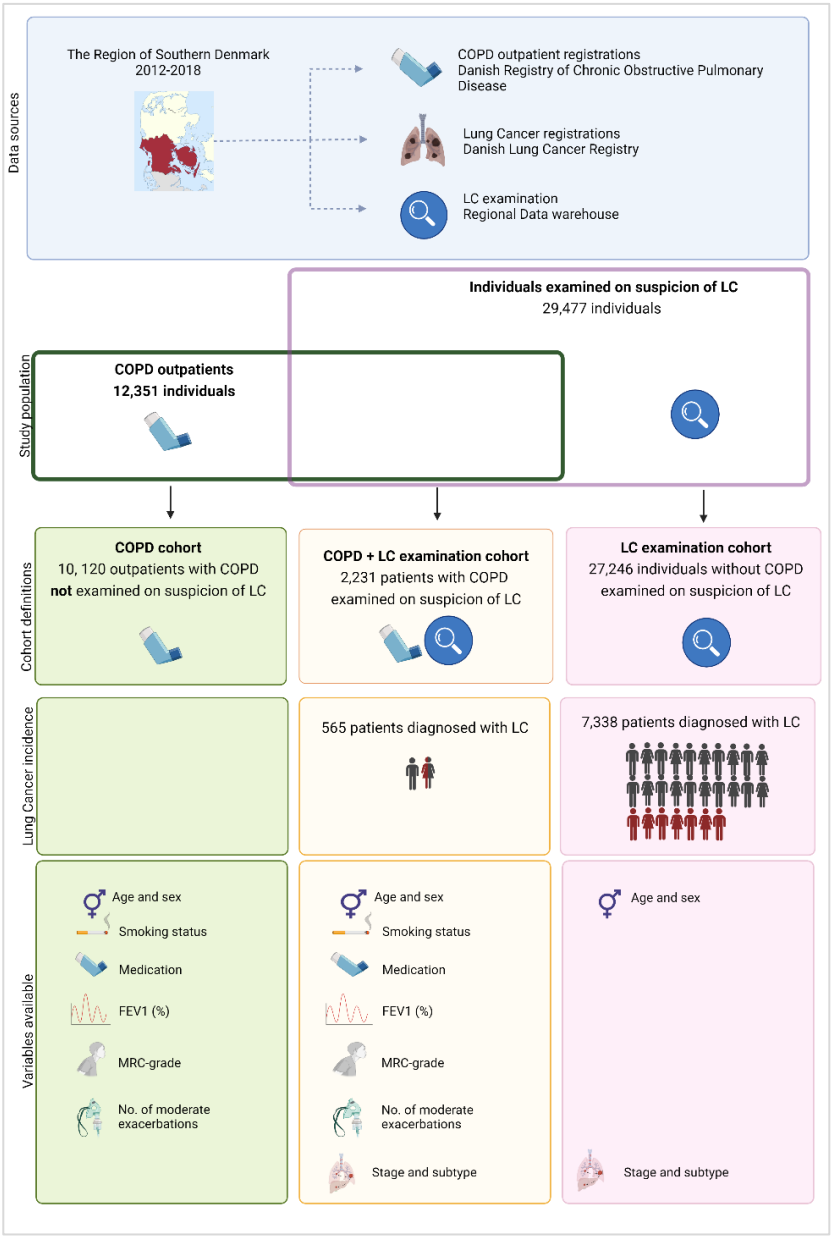


Figure 28: The three cohorts included in the study: COPD outpatients, patients examined in the LC fast-track clinics, and the overlapping cohort (COPD+LC examination cohort). Created with Biorender.com.

Data on outpatients with COPD were sourced from the DrCOPD managed by the Danish Clinical Quality Program [120]. This dataset includes individuals aged 30 and above and was limited to outpatient visits. COPD registrations from general practice were not yet accessible and therefore not included in the study. The population referred for LC examination was derived from the population in Article I, which included all individuals examined for LC during the study period. We combined COPD outpatients with those examined for suspected LC, labeling individuals present in both datasets as the "COPD + LC examination cohort". The remaining cohorts were designated as the "COPD cohort" and the "LC examination cohort". Information on diagnosis, stage, and subtype of LC was obtained from the Danish Lung Cancer Registry [113].

Variables

In the "COPD cohort," we included COPD-related variables from the most recent outpatient visit, as data completion rates were higher in recent years. For the "COPD + LC examination cohort," we used COPD-related variables registered closest to the LC examination, since our primary interest was data nearest to this event. In the "LC examination cohort," LC patients were registered with their date of diagnosis, while non-LC patients were registered with the date of their initial LC examination.

Smoking status was categorized into four groups: non-smoker, current smoker, former smoker, and unknown. Prescription medication data were collected in the following combinations: long-acting β-agonists (LABA), LABA+inhaled corticosteroids (ICS), long-acting muscarinic antagonists (LAMA), LABA+LAMA, LABA+LAMA+ICS and ICS. The variable “inhalation devices” covered all of the mentioned medicaments and was marked as present if any of them was prescribed.

Each variable was binary, with a positive outcome (1) indicating at least one prescription of inhalation medication during the specified year, while a negative outcome (0) signifying no collection within the year. Forced expiratory volume in 1 s (FEV_1), Medical Research Council (MRC) dyspnea score, and the frequency of moderate exacerbations were recorded as categorical variables (see Article V for further details).

Statistical analyses

Distributions were presented as fractions for categorical variables and medians with IQR for continuous variables. To test associations between groups, categorical variables were compared using the Chi-squared test, while continuous variables were analyzed with the Wilcoxon signed-rank test. A significance level of 0.01 was used in all tests to account for multiple comparisons. All statistical analyses were performed using STATA version 17.0.

Article V: Results

Cohort distributions & LC incidence

During the study period, 12,351 individuals were followed with COPD in outpatient clinics in the Region of Southern Denmark. During the same time 29,477 patients underwent evaluation for suspected LC of which 2,231 were both followed with COPD and examined on suspicion of LC. This overlapping group constituted 18% of all COPD outpatients and 8% of individuals examined for LC suspicion.

A total of 7,903 of the patients undergoing LC examination received a diagnosis of LC, among whom 565 were also diagnosed with COPD. This led to an LC incidence of 4.6% among all 12,351 COPD outpatients. The LC incidence among all patients examined on suspicion of LC was 26.8%, with no statistically significant difference between the COPD and non-COPD patients (25.3% vs. 26.9%, p=0.099).

Data on LC stage was available for 7,596 of the 7,903 LC patients. The LC patients with COPD included a higher proportion of stage I-II LC patients compared to the LC patients without COPD (46.2% vs. 26.3%) ($p<0.001$ for all) **(Figure 29)**. The overall distribution of pathological subtypes was comparable, although squamous cell carcinoma was represented to a higher degree among the LC patients with COPD compared to the group of LC patients without COPD (25.3% and 19.6%, respectively, $p<0.001$) (Article V, supplementary Table 4 and Figure 1).

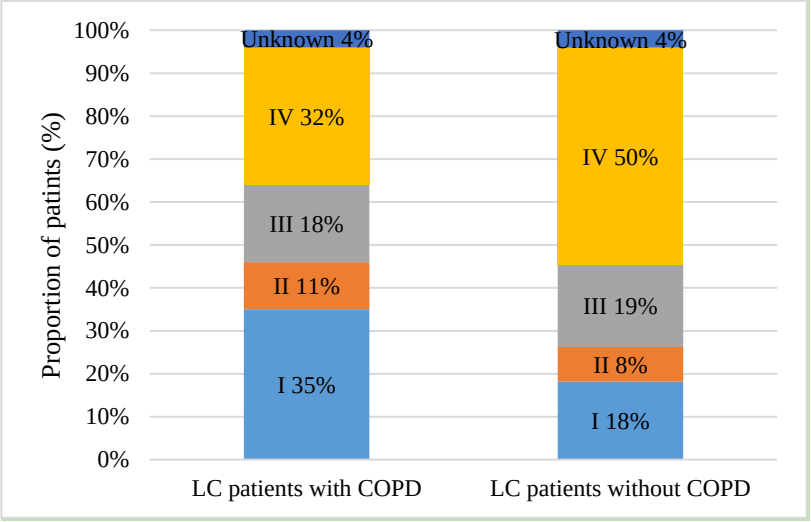


Figure 29: LC stage distribution in the cohort with and without COPD.

Comparison of the "COPD cohort" and the "COPD+LC examination cohort"

The COPD cohort was compared to the COPD+LC examination cohort in order to investigate what characterizes the patients referred for LC examination. Table 1 in Article V displays the comprehensive comparison of variables. Overall, the differences between the groups were small. However, key findings include that several medications were prescribed to a higher proportion of patients in the COPD+LC examination cohort, e.g. LABA+ICS, LAMA, and overall inhalation devices. Additionally, a larger proportion of patients in the COPD+LC examination cohort had an MRC grade >2 (69% vs. 66%, p=0.002) and experienced ≥2 exacerbations compared to the COPD cohort (30% vs. 24%, p<0.001).

07 Discussion

Summary of findings

The articles included in this thesis focus on populations in high or medium risk of LC and the development of prediction models for these cohorts. They offer detailed descriptions and characterizations of the populations (Articles I and V) and present prediction models using various methodologies (Articles II and III). Additionally, the aspect of data completeness and the most important combination of data to include in a prediction model is assessed (Article IV).

In Article I we observed an LC incidence of approximately 29% in the LC fast-track clinics, indicating that this population is particularly high-risk, primarily composed of current or former smokers. Surprisingly, most individuals did not have hospital-diagnosed comorbidities, but the fact that COPD medication was more prevalent in LC patients suggests that many mild to moderate COPD cases were managed in general practice. Despite relatively subtle changes in laboratory results, these continuous variables can capture more detailed information and nuances, which is relevant in robust prediction models. Symptoms such as weight loss, fatigue, and pain were more common in LC patients, while hemoptysis and fever were more prevalent in the non-LC group. These patterns remained consistent in an analysis of a sub-cohort with complete data.

In **Articles II** and **III** the high-risk cohort from Article I was used, focusing only on age, sex, smoking status, and laboratory results. Among the ML models developed in Article II, the DES model demonstrated the most robust performance, although it did not surpass the PLCOm2012 model or the Medial EarlySign model using the same types of data. The BN model performed similarly to the DES model but showed resilience when dealing with up to 30% missing data, which is a significant advantage in clinical settings.

In **Article IV** the optimal model was evaluated across data completeness and size as well as the best combination of variables. It was demonstrated that a model trained on high-quality data could still perform acceptably when applied to lower-quality data with higher rates of missing data across different datasets. Performance impact was minimal up to 39% missing data. Consistent with the findings in Articles II and III, age, smoking status, and laboratory data were the most significant predictors of LC risk.

In **Article V** we examined the demographics and LC incidence among patients with COPD in outpatient clinics. Surprisingly, nearly one-fifth of these patients were referred for LC diagnostics, with an LC incidence more than ten times higher than the background population. Additionally, LC patients with COPD were diagnosed at earlier stages

compared to those without COPD. The high prevalence of LC in this cohort, combined with their regular hospital visits, highlights the potential of considering this group for LC screening. In line with the focus of the present thesis, it also emphasizes the relevance of developing and validating LC prediction models for this population.

From model development to implementation

Figure 29 depicts the journey from algorithm development to implementation, emphasizing its lengthy nature. While initial steps like development and validation are widely practiced, relatively few algorithms advance to full implementation due to numerous challenges. This thesis concentrates primarily on model development, which will therefore be thoroughly examined in subsequent sections, contextualizing its strengths and limitations. The subsequent stages are also outlined to provide a comprehensive overview of the entire process, although discussed more briefly.

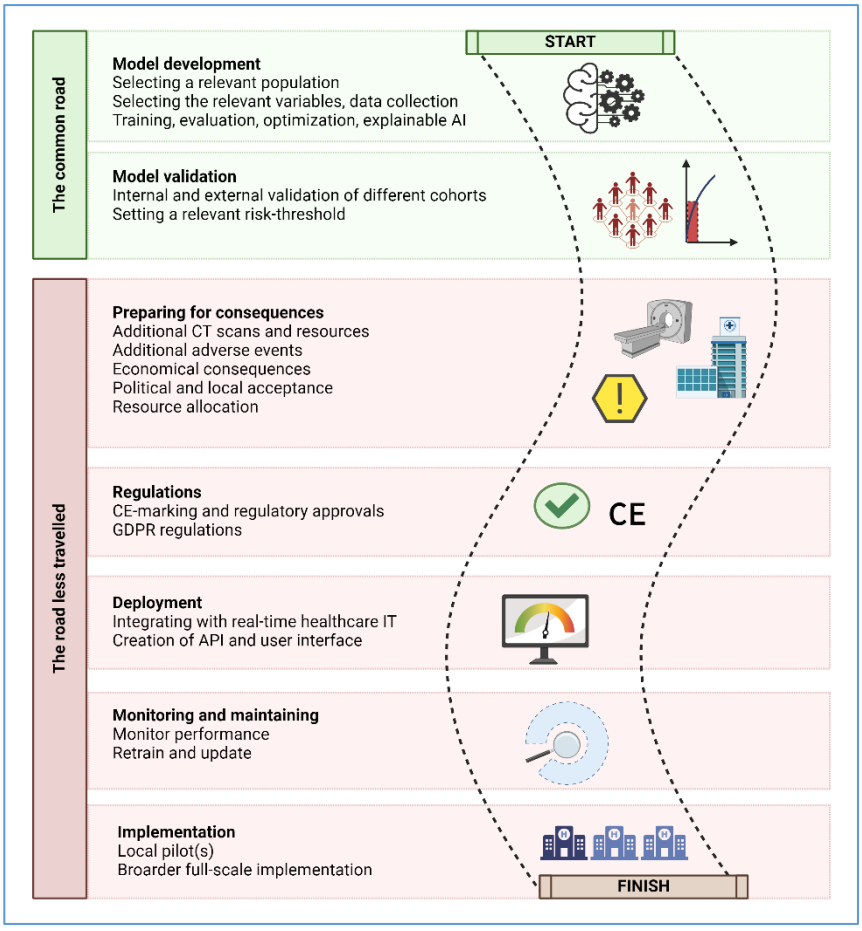


Figure 29: The journey from creating a prediction model to its implementation depicted as a road. The common road represents the development and validation of the model, while the less travelled road illustrates the steps involved in implementation. Created with Biorender.com. API: Application Programming Interface; CE: European Conformity; GDPR: General Data Protection Regulation. Created with Biorender.com.

The common road

Selecting a relevant population

When creating a prediction model, it is crucial to consider the population to be used, as the model's generalizability is often limited to the population it was developed on. This is especially important in relation to the incidence of the outcome of interest, in this case LC, where the LC incidence in the dataset used to train a model should reflect the population it will ultimately be applied to.

In this thesis, the models were built using a high-risk cohort of patients who were referred to the LC fast-track clinics due to symptoms or findings on CT or X-ray, with an overall LC incidence of approximately 25%. Given this high LC incidence, with a number-needed-to-screen of only 4, the clinical utility of such a model may be questionable. At the time of study planning, we were aware of the issues in using this population. However, local experience with using clinical data for prediction models was limited, and access to broader data from general practice was unavailable. This constrained us to using hospital data, and we chose patients from the LC fast-track unit because cases as well as controls had representative information on smoking status and laboratory analyses performed at the time of examination. Furthermore, we found it relevant that cases and controls represented equal “times of suspicion," challenging the model's discrimination. Other studies often matched controls to LC cases, potentially resulting in a healthier control population and easing model discrimination [64, 121]. Therefore, the benefit of having enough LC cases to create a prediction model and sufficient variable information on cases and controls outweighed the limitations at the time of study planning and were the main reasons for the selected study cohorts.

After constructing the models on the high-risk population, we aimed for a new population more suitable for LC screening while maintaining data collection capability. We explored the COPD outpatient cohort and found an LC incidence of approximately 5%, which is more than ten times that of the background population [65]. Although this seems relatively high, it only indicates the overlap between the two cohorts and not an in-depth analysis of the order of events, which is a limitation of Study V.

In about half the cases, the LC diagnosis was registered before the first visit to the COPD outpatient clinic in the study period, and in the remaining cases, it was the other way around. Since outpatient visits were not marked as initial or follow-up, we would only have been able to identify the first visit by collecting COPD visits in the years prior to inclusion.

This was not done, so the true rate of LC diagnosis among COPD patients is unknown and may be lower than 5%. Establishing a period for follow-up after the end of the study to ensure controls were truly controls and not diagnosed shortly after the study ended was not done but will be considered in future studies.

None of these limitations change the fact that the COPD outpatient cohort is promising for future LC screening models. The LC incidence of 3-5% and the regular hospital visits by these patients call for further investigation, since adherence remains one of the biggest challenges.

Generalizability to other populations is crucial not only in terms of LC incidence but also the distribution of other variables. The USPSTF LC screening criteria have been criticized to favor non-Hispanic whites since the criteria were derived from the NLST trial performed primarily on this demography [122]. The consequence is under-screening of high-risk individuals in minority populations. In our study the patients from the LC fast-track clinic were generally high-risk with age and smoking history fitting the relevant screening population. Similarly, the COPD outpatients were comparable in demographics, representing a relevant at-risk cohort.

Selecting relevant variables

In this thesis, the prediction models developed in articles II and III included the variables age, sex, binary smoking status, and 20 regularly analyzed blood biomarkers. The results of the SHAP-analyses of article II revealed that a status of former or active smoker together with high age were the factors mostly attributing to the predicted risk. High calcium, LDH and neutrophil counts were the three most influential biomarkers.

Smoking status

Smoking status was obtained by manually annotating free-text from EHR data. The initial goal was to gather detailed information on smoking duration and intensity, measured in pack-years and years since quitting, to compare screening criteria with other guidelines. However, this level of detail was only available for a small subset of individuals, mostly LC patients. Consequently, smoking status was annotated simply as former, active, or never-smoker, which is a general limitation of our studies.

The overall lack of structured smoking data in Danish healthcare led to a spin-off project in collaboration with A. Ebrahimi from SDU Data Science. The annotated smoking data were used as labels to develop a natural language processing model capable of automatically predicting smoking status in binary format from EHR data [4]. Future projects will include

validation of this model and the development of an advanced model to include information on pack-years when available.

Even with structured data, recall bias and social desirability bias are still relevant concerns, as individuals may underreport smoking intensity to avoid negative judgment [123]. This makes the work on DNA methylation particularly interesting, as several genes are strongly associated with smoking [124, 125]. DNA methylation biomarkers can provide an objective measure of smoking behaviour and exposure, which, along with LC-specific biomarkers, has the potential to improve LC screening and prediction accuracy. These methods, however, require invasive testing and cannot rely on previous clinical or laboratory assessments, which are more commonly available. Therefore, it remains relevant and interesting to use standard laboratory analyses as evaluated in this thesis.

Age

Advanced age is a well-established risk factor for all cancers, as the accumulation of mutations increases with age. However, in the realm of risk modelling, age is also recognized as a source of bias. A significant argument against using risk models is that incorporating age skews eligibility towards older individuals, who are at the highest risk according to these models. Although older individuals have a higher risk for LC, they also have fewer potential life-years saved by screening. Thus, using risk-based criteria may increase the number of LC deaths averted but not necessarily the number of life-years gained [103, 126]. Apart from our DES-model, age was also a top influential factor in the PLCOm2012 and the Medial EarlySign model. When broadly applying the NLST and NELSON criteria to our high-risk cohort, we estimated that only 54% would be eligible for screening, primarily due to individuals exceeding the upper age limit of 74 years. Given these considerations, it may be relevant to incorporate life-years gained into future risk modeling.

Laboratory results

The distribution of laboratory results assessed in Articles II and III proved to be useful in LC prediction, with Study II highlighting high levels of LDH, calcium and sodium the most influential in the process. Although median levels often differed significantly between the LC and non-LC cohorts, the differences were not clinically significant and most values remained within the standard reference interval. Nevertheless, the DES model achieved its maximum F1-score with the top 10 features and only a 2% decrease when reducing to the top 5 features. When the model was reduced to only include the top two features of smoking history and LDH, the F1-score dropped by an additional 2%. This aligns with the findings of Article IV, where BNs incorporating laboratory data outperformed those using

only age, sex, and smoking history. Hence, including laboratory results can enhance the prediction model compared to a model that only considers age and smoking status as seen in the current USPSTF recommendations.

In this study, we focused on laboratory analyses from the LC fast-track units, where all examined patients are expected to undergo similar tests. Had we included laboratory analyses from the general population, selection bias would likely have occurred due to physicians primarily ordering tests for patients with symptoms or comorbidities. As a result, asymptomatic individuals would be less likely to have blood samples taken, which could skew a prediction model towards patients with more symptoms and comorbidities. To generalize to a broader population, the model must be developed using a sufficiently diverse dataset that covers the entire population, ensuring a balanced distribution of missing information between LC and non-LC patients.

Comparing different categories of data in article IV, laboratory results turned out to be the most influential; perhaps because they are continuous variables, which allow for more precise and nuanced analyses although discretized before using the BN approach. The discretization was based on the optimal split in the data distributions and appeared to have greater discriminative value compared to dichotomized variables such as comorbidity measures. A significant limitation in Study IV was that we only compared groups of data (comorbidity, laboratory results, smoking, and symptoms) and not specific variable combinations. It is possible that specific combinations such as a diagnosis of COPD with certain laboratory analyses and COPD prescription medications, might enhance prediction. To better reflect this matter, future analyses will consider separating all variables rather than pooling them into categories.

Comorbidity and medication data

In addition to the limitation of categorizing all comorbidity data (along with medication and GP data) into one dataset, the lack of significance might also stem from the fact that these data were recorded solely at the hospital level and not collected from general practice. Although specific types of medication were used as proxies for diseases such as mild/moderate hypertension, COPD, and diabetes, this approach was not entirely accurate and introduced bias. Nevertheless, the results of article I revealed that more than half of both LC and non-LC patients were not registered with any of the included comorbidities at the hospital level. This finding is notable because it challenges the common perception of an LC patient as typically having COPD or cardiovascular disease, indicating that other factors beyond hospital-level comorbidities must be considered. Including IPCP codes to

reflect diagnoses registered in general practice would be an interesting path for future research.

Symptoms

In Article IV we found that symptoms were among the least impactful data set. This is an important aspect since symptoms data are not commonly collected and difficult to obtain. Notably, hemoptysis, the most well-known "red-flag" symptom of LC, was the most common symptom in both the LC and the non-LC cohorts, although it appeared more frequently in the non-LC group. This could be due to clinical practice commonly referring patients for potential bronchoscopy based on prolonged hemoptysis, making it the most common referral symptom in the LC fast-tracks. Consequently, the representation of symptoms cannot be generalized to the broader population eligible for screening. While the distribution of symptoms among LC patients might not change significantly, the frequency of symptoms in non-LC patients would likely be much lower than observed in the study. If this hypothesis holds true, the difference in symptom frequency between LC and non-LC individuals would be more pronounced, potentially enhancing the clinical utility of symptoms in a prediction model.

Optimal data requirements

In addition to the mentioned data sources, several relevant ones were unavailable for collection, such as radon exposure, air pollution, symptoms or diagnoses from general practice, and occupational exposures. To develop prediction models capable of identifying the 10-15% of LC patients with no smoking history, gathering some of these risk factors would be essential. Future efforts could gain insight into these factors through linkage with Statistics Denmark and access to general practice data, although certain exposures may remain challenging to obtain.

The optimal data requirements in a prediction model can be summarized as follows. The data should be clinically relevant, linked through either causation (carcinogenesis) or associated to LC and be relatively easy to obtain, both in real-time and from previous results. Data capture and potential tests should be inexpensive, ideally non-invasive, and carry a low risk of adverse events. Additionally, it should be feasible to gather the data "behind the scenes" without requiring patients to undergo specific tests or complete questionnaires about habits for a predicted risk to be calculated.

Of the various types of data examined in this thesis, regular laboratory analysis data meet these requirements. If an automated NLP algorithm for collection of detailed smoking status from patient files becomes applicable, the combination of smoking data, age, sex, and

laboratory results provides the listed benefits. This would potentially outweigh the performance of using biomarkers, or at least present a relevant alternative.

Model evaluation

Criteria for an optimal prediction model, in terms of performance evaluation, include demonstrating strong discrimination, calibration (particularly within its intended interval of use), and clinical utility. Moreover, the model should effectively predict the onset of LC before conventional diagnostics, by detecting earlier stages of the disease.

The models developed in this thesis all performed inferiorly to the PLCO and Medial EarlySign models. However, as discussed in article II, the mentioned studies differ significantly in population, selection of controls, prediction intervals, included variables, and statistical methods applied. For a direct comparison to be meaningful, all these criteria would need to be aligned. This is a common challenge in the comparison of models, which makes it valuable that articles II and III directly compare performances on the same population using the exact same data input. While the BN models developed in study III performed slightly inferiorly to the DES model in study II, they were comparable in terms of discrimination and clinical utility. A notable feature of the BN model is its robust handling of a large amount of missing data.

A direct comparison using the DES model with varying rates of missing variables was not performed, but this would be an interesting future comparison. The SHAP analysis is a significant advantage of the DES model, providing great opportunities to offer individual insight into clinicians and patients. In contrast, DAGs did not offer the same level of clinically relevant insight on an individual level, especially when many variables were included, making them more comprehensive.

In the US, the nationally recommended yearly screening of eligible individuals is primarily based on multi-year risk assessments, typically 5 or 6-year risk estimates. LC often develops over several years, so long-term risk assessments are more effective at capturing the cumulative risk and identifying individuals who may develop LC over time. Given that LC is relatively rare within a one-year window, longer horizons will capture more events, thereby improving the statistical power and robustness of the models. The drawbacks of models based on a long risk horizon include ethical considerations, e.g. increased psychological consequences, and a higher risk of adverse events with annual screenings.

The prediction models developed in this thesis aim to estimate the risk of LC based on information available at the time of examination or diagnosis. Thus, the risk is assessed at time zero, making it more of an LC detection model rather than a true prediction model.

This is a notable shortcoming in our studies, which we tried to address by incorporating previous laboratory analyses to create a model capable of estimating the risk of LC, for example six months before the diagnosis. The approach was not feasible due to limited laboratory analyses performed in the previous period. New investigations are ongoing, examining previous laboratory results of high-risk patients and those followed in the COPD outpatient clinic.

Model validation

An optimal screening model should undergo both internal and external validation across diverse cohorts, potentially including different nationalities or at least match the target population to the extent possible. The risk threshold of whether an individual should be screened must balance predictive accuracy with the considerations on consequences discussed in the next section.

This thesis has not yet externally validated any of the developed algorithms. We are currently collaborating with the creators of the Medial EarlySign models to conduct external validation on both our high-risk cohort and the COPD outpatient cohort described in Article V. While methodological differences exist, this effort promises valuable insight into comparing model performances.

The road less travelled

Preparing for consequences

Before implementing a well-validated prediction model, a comprehensive analysis of the ensuing consequences is essential. Initially, the chosen threshold must align with clinically relevant performance metrics and consider the number-needed-to-screen within the existing screening capacity. Studies evaluating the cost-effectiveness of LC screening trials consistently highlight the importance of identifying high-risk individuals, with CT scans being the primary cost factor [127]. The capacity of CT scanning facilities and the shortage of radiologists also pose significant considerations.

One potential solution involves integrating AI-based tools into CT interpretations. The EU-funded multinational 4-IN-THE-LUNG-RUN project incorporates AI-based CT-readings, demonstrating a 79% agreement in nodule categorization between AI and human reads

[128]. This could substantially reduce radiologist workload and enhance consistency across different institutions.

Increased CT scanning may call for more surgical capacity. Studies from the US suggest that a 50% adherence rate to LC screening could necessitate a 37% increase in surgical capacity. Given the importance of targeting high-risk individuals for cost-effective implementation, some argue for raising risk thresholds, e.g. to 2.5% using models like PLCOm2012 [127].

Critical to the success of any screening program is a robust recruitment and communication strategy aimed at reaching high-risk individuals. This remains a significant challenge and is underscored by the low adherence rates observed in US national screening efforts. It is essential to address these complexities to optimize the effectiveness and efficiency of LC screening initiatives. Recruiting at-risk patients at a hospital level, e.g. COPD outpatients, is one approach suggested in this thesis. Community outreach or online invitations is another method. While inspiration can be drawn from global approaches, recruitment strategies must align with the unique healthcare system, data infrastructure, and high-risk demographics of the country where they are applied. Hence, it is essential to design a recruitment strategy tailored specifically to the Danish high-risk population and to evaluate its efficacy and feasibility in pilot studies before implementing it at a national level.

Regulations and deployment

When it comes to medical AI, obtaining regulatory approval to sell and distribute products within the EU is a complex and evolving process that varies across countries. In March 2024, the European Parliament passed the groundbreaking AI Act, marking the EU's first comprehensive regulation addressing the use of AI. Meeting the regulations as well as obtaining CE marking involves a rigorous and lengthy process, and the number of Danish AI medical startups that have achieved certification is limited.

During the deployment phase of a medical AI prediction tool, proper installation and configuration is essential to integrate with existing healthcare systems. The tool must utilize real-time data and be able to aggregate information from previous assessments. Collaborative development of user interfaces and training programs with clinicians optimizes synergy and ensures the AI enhances clinical processes efficiently with minimal disruption [129].

Monitoring and maintaining and implementation

In the monitoring and maintenance phase, continuous monitoring of model performance through systematic logs is crucial to detect potential drops due to factors like data drift. Changes in patient demographics, disease incidence, healthcare service setups (e.g. during the COVID-19 pandemic), and modifications in diagnostic procedures require the AI model to be retrained on new data to ensuring adaptability and resilience to evolving data distributions [130].

Prior to full-scale implementation, smaller pilot studies are usually conducted in controlled environments such as specific departments or healthcare facilities. They serve to validate the AI solution, gather performance data, assess feasibility, and identify potential areas for improvement before broader implementation, where regulatory oversight can be more challenging. Once proven effective in pilot studies, plans can be made to scale up deployment across multiple facilities or regions. This includes ensuring the capability of the IT infrastructure to support full-scale implementation and establish robust data governance practices. Data governance encompasses aspects such as security, quality, access and overall data accountability as described in the FAIR principles for data management [131].

08 Conclusion and perspectives

The potential of AI in early detection of LC was investigated in the thesis focusing primarily on high and medium-risk populations.

In the high-risk LC fast-track population, the LC incidence was approximately 29% and consisted predominantly of current and former smokers. Although the majority lacked a hospital-registered diagnosis, the rate of COPD medication usage was higher in LC patients. Together with smoking status and age the subtle changes in laboratory parameters were able to detect LC to a moderate degree using a DES as well as a BN approach. The BN model's resilience to missing data made it perform well even with up to 30% missing data.

When expanding the BN models to include other types of data it was confirmed that age, smoking status, and laboratory results were the most influential datasets in detecting LC. Furthermore, a model trained on a small, detailed dataset showed robustness in the validation of a larger dataset with more missing data, resembling real-world screening scenarios where data may be incomplete.

Investigation of the medium-risk population (COPD outpatients) revealed an LC incidence of approximately 5%, which is more than ten times higher than the background population. Nearly one-fifth of the COPD patients were referred for LC diagnostics, and LC patients with COPD were diagnosed at earlier stages compared to LC patients without COPD.

Overall, the research presented in this thesis highlights the importance of refining LC prediction models and explores the potential of using EHR-based data to detect LC. The diverse array of EHR-based data that were gathered and transformed demonstrated the feasibility of using these data sources to develop prediction models for high and medium-risk populations. Although the developed models did not surpass established prediction models such as PLCOm2012 or Medial EarlySign, direct comparison was challenged by discrepancies in study design and demographics.

Ongoing research includes studies to validate the Medial EarlySign model on our high-risk cohort and the COPD outpatient cohort. They aim to make direct comparisons and evaluate whether an early detection model can be effective in a Danish context, where rates and completeness of laboratory data may differ from those in other countries where data collection is more routinely performed.

The present thesis underscores the importance of considering COPD outpatients in the screening for LC. This aligns with future prospective studies where prediction models and

potential biomarkers will be tested in COPD patients undergoing regular visits to the outpatient clinic.

The potential of EHR-based data in LC prediction should be seen as a valuable supplementary tool rather than a substitute for effective biomarkers such as circulating tumor DNA. Alongside advances in AI and biomarker technology, smoking cessation must remain a key focus, especially in younger generations. Looking 10-15 years ahead, the ideal scenario would involve a significant decline in smoking rates in Denmark along with the successful implementation of LC screening supported by a recruitment strategy targeting at-risk populations. This approach would depend on accurate biomarkers, potentially supplemented or enhanced by EHR data, especially in situations where biomarkers are difficult to obtain. The vision also encompasses comprehensive nationwide data on smoking status and other risk factors, with AI well-integrated into the analysis of CT scans. By continuing to prioritize research and implementation in these areas, we can aspire to a future with significantly reduced LC incidence and mortality, ultimately benefiting countless patients.

09 References

1. 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.* AME Publishing Company; 2023; 12: 2392.

2. Flyckt RNH, Sjodsholm L, Henriksen MHB, Brasen CL, Ebrahimi A, Hilberg O, Hansen TF, Wiil UK, Jensen LH, Peimankar A. Pulmonologists-Level lung cancer detection based on standard blood test results and smoking status using an explainable machine learning approach. *arXiv Prepr. arXiv2402.09596* 2024; .

3. Bang Henriksen M, Hansen TF, Jensen LH, Brasen CL, Borg M, Hilberg O, Løkke A. Lung cancer among outpatients with COPD: a 7-year cohort study. *ERJ Open Res.* 2024; 10: 64–2024.

4. Ebrahimi A, Henriksen MBH, Brasen CL, Hilberg O, Hansen TF, Jensen LH, Peimankar A, Wiil UK. Identification of patients' smoking status using an explainable AI approach: a Danish electronic health records case study. *BMC Med. Res. Methodol.* 2024; 24: 114.

5. World Health Organization. Lung cancer [Internet]. 2023 [cited 2023 Dec 13].Available from: https://www.who.int/news-room/fact-sheets/detail/lung-cancer.

6. Thandra KC, Barsouk A, Saginala K, Aluru JS, Barsouk A. Epidemiology of lung cancer. *Contemp. Oncol. (Poznan, Poland)* Poland; 2021; 25: 45–52.

7. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* Wiley Online Library; 2021; .

8. World Cancer Research Fund International. Lung cancer statistics [Internet]. 2022 [cited 2023 Dec 13].Available from: https://www.wcrf.org/cancer-trends/lung-cancer-statistics/.

9. Christensen NL, Jekunen A, Heinonen S, Dalton SO, Rasmussen TR. Lung cancer guidelines in Sweden, Denmark, Norway and Finland: a comparison. *Acta Oncol.* Sweden; 2017; 56: 943–948.

10. National cancer institute. Cancer Stat Facts: Lung and Bronchus Cancer [Internet]. 2024 [cited 2024 May 1].Available from: https://seer.cancer.gov/statfacts/html/lungb.html.

11. Polanco D, Pinilla L, Gracia-Lavedan E, Mas A, Bertran S, Fierro G, Seminario A, Gómez S, Barbé F. Prognostic value of symptoms at lung cancer diagnosis: a three-

year observational study. *J. Thorac. Dis.* China; 2021; 13: 1485–1494.

12. Yi M, Jiao D, Qin S, Chu Q, Wu K, Li A. Synergistic effect of immune checkpoint blockade and anti-angiogenesis in cancer treatment. *Mol. Cancer* Springer; 2019; 18: 1–12.

13. Rudin CM, Avila-Tang E, Harris CC, Herman JG, Hirsch FR, Pao W, Schwartz AG, Vahakangas KH, Samet JM. Lung cancer in never smokers: molecular profiles and therapeutic implications. *Clin. Cancer Res.* AACR; 2009; 15: 5646–5661.

14. Hecht SS. Lung carcinogenesis by tobacco smoke. *Int. J. cancer* Wiley Online Library; 2012; 131: 2724–2732.

15. Hecht SS. Tobacco smoke carcinogens and lung cancer. *J. Natl. cancer Inst.* Oxford University Press; 1999; 91: 1194–1210.

16. Diaz M, Garcia M, Vidal C, Santiago A, Gnutti G, Gómez D, Trapero-Bertran M, Fu M, research group LCPL, others. Health and economic impact at a population level of both primary and secondary preventive lung cancer interventions: a model-based cost-effectiveness analysis. *Lung Cancer* Elsevier; 2021; 159: 153–161.

17. Laconi E, Marongiu F, DeGregori J. Cancer as a disease of old age: changing mutational and microenvironmental landscapes. *Br. J. Cancer* Nature Publishing Group UK London; 2020; 122: 943–952.

18. Pikor LA, Ramnarine VR, Lam S, Lam WL. Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung cancer* Elsevier; 2013; 82: 179–189.

19. Danish Lung Cancer Group. Clinical guideline [Internet]. 2020 [cited 2022 Nov 19].Available from: https://www.lungecancer.dk/wp-content/uploads/2020/12/DLCG_visitation_diagn_stadie_AdmGodk141220.pdf.

20. Tammemagi CM, Neslund-Dudas C, Simoff M, Kvale P. Impact of comorbidity on lung cancer survival. *Int. J. cancer* Wiley Online Library; 2003; 103: 792–802.

21. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin. Epidemiol.* Taylor \& Francis; 2015; : 449–490.

22. McDonald L, Carroll R, Harish A, Tanna N, Mehmud F, Alikhan R, Ramagopalan S V. Suspected cancer symptoms and blood test results in primary care before a diagnosis of lung cancer: a case--control study. *Futur. Oncol.* Future Medicine; 2019; 15: 3755–3762.

23. Shiels MS, Pfeiffer RM, Hildesheim A, Engels EA, Kemp TJ, Park J-H, Katki HA, Koshiol J, Shelton G, Caporaso NE, others. Circulating inflammation markers and prospective risk for lung cancer. *J. Natl. Cancer Inst.* Oxford University Press US;

2013; 105: 1871–1880.

24. Forkasiewicz A, Dorociak M, Stach K, Szelachowski P, Tabola R, Augoff K. The usefulness of lactate dehydrogenase measurements in current oncological practice. *Cell. \& Mol. Biol. Lett.* BioMed Central; 2020; 25: 1–14.

25. Acharya S, Kale J, Rai P, Anehosur V, Hallikeri K. Serum alkaline phosphatase in oral squamous cell carcinoma and its association with clinicopathological characteristics. *South Asian J. cancer* Thieme Medical and Scientific Publishers Pvt. Ltd.; 2017; 6: 125–128.

26. Olesen F, Hansen RP, Vedsted P. Delay in diagnosis: the experience in Denmark. *Br. J. Cancer* England; 2009; 101 Suppl: S5-8.

27. Danish Lung Cancer Group. Clinical guideline [Internet]. 2023Available from: https://www.dmcg.dk/siteassets/kliniske-retningslinjer---skabeloner-og-vejledninger/kliniske-retningslinjer-opdelt-pa-dmcg/lungecancer/dlcg_visitation_diagn_stadie_v.3.0_admgodk_121223.pdf.

28. The Danish Ministry of Health. "Jo før jo bedre" Tidlig diagnostik, bedre behandling og flere gode leveår for alle [Internet]. 2014 [cited 2024 May 3].Available from: https://sum.dk/Media/637643658255283608/Jo før - jo bedre.pdf.

29. Steding-Jessen M, Engberg H, Jakobsen E, Rasmussen TR, Møller H. Progress against lung cancer, Denmark, 2008-2022. *Acta Oncol.* 2024; 63: 339–342.

30. Team NLSTR. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* Mass Medical Soc; 2011; 365: 395–409.

31. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, Lammers J-WJ, Weenink C, Yousaf-Khan U, Horeweg N, van 't Westeinde S, Prokop M, Mali WP, Mohamed Hoesein FAA, van Ooijen PMA, Aerts JGJ V, den Bakker MA, Thunnissen E, Verschakelen J, Vliegenthart R, Walter JE, Ten Haaf K, Groen HJM, Oudkerk M. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N. Engl. J. Med.* United States; 2020; 382: 503–513.

32. Aberle D, Adams A, Berg C, Black W, Clapp J, Fagerstrom R, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening New England Journal of Medicine 365 (5): 395-409 DOI 10.1056. *NEJMoa1102873* 2011; .

33. Field JK, Vulkan D, Davies MPA, Baldwin DR, Brain KE, Devaraj A, Eisen T, Gosney J, Green BA, Holemans JA, others. Lung cancer mortality reduction by LDCT screening: UKLS randomised trial results and international meta-analysis. *lancet Reg. Heal.* Elsevier; 2021; 10.

34. Paci E, Puliti D, Pegna AL, Carrozzi L, Picozzi G, Falaschi F, Pistelli F, Aquilini F, Ocello C, Zappa M, others. Mortality, survival and incidence rates in the ITALUNG randomised lung cancer screening trial. *Thorax* BMJ Publishing Group Ltd; 2017; 72: 825–831.

35. McWilliams AM, Mayo JR, Im Ahn M, MacDonald SLS, Lam SC. Lung cancer screening using multi-slice thin-section computed tomography and autofluorescence bronchoscopy. *J. Thorac. Oncol.* Elsevier; 2006; 1: 61–68.

36. dos Santos RS, Franceschini JP, Chate RC, Ghefter MC, Kay F, Trajano ALC, Pereira JR, Succi JE, Fernando HC, Júnior RS. Do current lung cancer screening guidelines apply for populations with high prevalence of granulomatous disease? Results from the First Brazilian Lung Cancer Screening Trial (BRELT1). *Ann. Thorac. Surg.* Elsevier; 2016; 101: 481–488.

37. Blanchon T, Bréchot J-M, Grenier PA, Ferretti GR, Lemarié E, Milleron B, Chagué D, Laurent F, Martinet Y, Beigelman-Aubry C, others. Baseline results of the Depiscan study: a French randomized pilot trial of lung cancer screening comparing low dose CT scan (LDCT) and chest X-ray (CXR). *Lung cancer* Elsevier; 2007; 58: 50–58.

38. Becker N, Motsch E, Gross M-L, Eigentopf A, Heussel CP, Dienemann H, Schnabel PA, Eichinger M, Optazaite D-E, Puderbach M, others. Randomized study on early detection of lung cancer with MSCT in Germany: results of the first 3 years of follow-up after randomization. *J. Thorac. Oncol.* Elsevier; 2015; 10: 890–896.

39. Kreftforeningen. Lung Cancer Screenings to begin in Norway [Internet]. [cited 2024 Jun 27].Available from: https://kreftforeningen.no/en/lung-cancer-screenings-to-begin-in-norway/.

40. Kaneko M, Eguchi K, Ohmatsu H, Kakinuma R, Naruke T, Suemasu K, Moriyama N. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. *Radiology* 1996; 201: 798–802.

41. Zhao S-J, Wu N. Early detection of lung cancer: Low-dose computed tomography screening in C hina. *Thorac. Cancer* Wiley Online Library; 2015; 6: 385–389.

42. Hu P, Dai M, Shi J, Ren J, Li J, Liao X, Du L, Liu Y, Chen Z, Wu N, others. The feasibility study of a randomized cancer screening trial in China. *Cancer Res.* AACR; 2016; 76: 1795.

43. Wille MMW, Dirksen A, Ashraf H, Saghir Z, Bach KS, Brodersen J, Clementsen PF, Hansen H, Larsen KR, Mortensen J, others. Results of the randomized Danish lung cancer screening trial with focus on high-risk profiling. *Am. J. Respir. Crit. Care Med.* American Thoracic Society; 2016; 193: 542–551.

44. Pinsky PF. Lung cancer screening with low-dose CT: a world-wide view. *Transl. lung cancer Res.* AME Publications; 2018; 7: 234.

45. van der Aalst C, Vonder M, Hubert J, Moldovanu D, Schmitz A, Delorme S, Kaaks R, ten Haaf K, Oudkerk M, de Koning H. P1. 14-04 European lung cancer screening implementation: 4-IN-THE-LUNG-RUN trial. *J. Thorac. Oncol.* Elsevier; 2023; 18: S217.

46. U.S. Preventive Services Task Force. Final Recommendation Statement. Lung Cancer: Screening [Internet]. 2021 [cited 2024 May 22].Available from: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/lung-cancer-screening.

47. Grover H, King W, Bhattarai N, Moloney E, Sharp L, Fuller L. Systematic review of the cost-effectiveness of screening for lung cancer with low dose computed tomography. *Lung Cancer* Elsevier; 2022; 170: 20–33.

48. Wait S, Alvarez-Rosete A, Osama T, Bancroft D, Cornelissen R, Marušić A, Garrido P, Adamek M, van Meerbeeck J, Snoeckx A, others. Implementing lung cancer screening in Europe: taking a systems approach. *JTO Clin. Res. Reports* Elsevier; 2022; 3: 100329.

49. Martini K, Chassagnon G, Frauenfelder T, Revel M-P. Ongoing challenges in implementation of lung cancer screening. *Transl. Lung Cancer Res.* AME Publications; 2021; 10: 2347.

50. Ten Haaf K, van der Aalst CM, de Koning HJ, Kaaks R, Tammemägi MC. Personalising lung cancer screening: An overview of risk-stratification opportunities and challenges. *Int. J. cancer* Wiley Online Library; 2021; 149: 250–263.

51. Marcus PM, Lenz S, Sammons D, Black W, Garg K. Recruitment methods employed in the National Lung Screening Trial. *J. Med. Screen.* England; 2012; 19: 94–102.

52. van Iersel CA, de Koning HJ, Draisma G, Mali WPTM, Scholten ET, Nackaerts K, Prokop M, Habbema JDF, Oudkerk M, van Klaveren RJ. 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* United States; 2007; 120: 868–874.

53. Pinsky PF, Berg CD. Applying the National Lung Screening Trial eligibility criteria to the US population: what percent of the population and of incident lung cancers would be covered? *J. Med. Screen.* SAGE Publications Sage UK: London, England; 2012; 19: 154–156.

54. Røe OD, Markaki M, Tsamardinos I, Lagani V, Nguyen OTD, Pedersen JH, Saghir

Z, Ashraf HG. “Reduced” HUNT model outperforms NLST and NELSON study criteria in predicting lung cancer in the Danish screening trial. *BMJ open Respir. Res.* England; 2019; 6: e000512.

55. Dubin S, Griffin D. Lung cancer in non-smokers. *Mo. Med.* Missouri State Medical Association; 2020; 117: 375.

56. Aldrich MC, Mercaldo SF, Sandler KL, Blot WJ, Grogan EL, Blume JD. Evaluation of USPSTF lung cancer screening guidelines among African American adult smokers. *JAMA Oncol.* American Medical Association; 2019; 5: 1318–1324.

57. Pasquinelli MM, Tammemägi MC, Kovitz KL, Durham ML, Deliu Z, Rygalski K, Liu L, Koshy M, Finn P, Feldman LE. Risk Prediction Model Versus United States Preventive Services Task Force Lung Cancer Screening Eligibility Criteria: Reducing Race Disparities. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* United States; 2020; 15: 1738–1747.

58. Pasquinelli MM, Tammemägi MC, Kovitz KL, Durham ML, Deliu Z, Guzman A, Rygalski K, Liu L, Koshy M, Finn P, Feldman LE. Addressing Sex Disparities in Lung Cancer Screening Eligibility: USPSTF vs PLCOm2012 Criteria. *Chest* United States; 2022; 161: 248–256.

59. Marmor HN, Zorn JT, Deppen SA, Massion PP, Grogan EL. Biomarkers in lung cancer screening: a narrative review. *Curr. challenges Thorac. Surg.* NIH Public Access; 2023; 5.

60. Li Z, Shu J, Yang B, Zhang Z, Huang J, Chen Y. Emerging non-invasive detection methodologies for lung cancer. *Oncol. Lett.* Spandidos Publications; 2020; 19: 3389–3399.

61. Kammer MN, Lakhani DA, Balar AB, Antic SL, Kussrow AK, Webster RL, Mahapatra S, Barad U, Shah C, Atwater T, others. Integrated biomarkers for the management of indeterminate pulmonary nodules. *Am. J. Respir. Crit. Care Med.* American Thoracic Society; 2021; 204: 1306–1316.

62. Baldwin DR, Callister ME, Crosbie PA, O’Dowd EL, Rintoul RC, Robbins HA, Steele RJC. Biomarkers in lung cancer screening: the importance of study design. Eur. Respir. J. Eur Respiratory Soc; 2021.

63. Wang X, Zhang Y, Hao S, Zheng L, Liao J, Ye C, Xia M, Wang O, Liu M, Weng CH, Duong SQ, Jin B, Alfreds ST, Stearns F, Kanov L, Sylvester KG, Widen E, McElhinney DB, Ling XB. Prediction of the 1-Year Risk of Incident Lung Cancer: Prospective Study Using Electronic Health Records from the State of Maine. *J. Med. Internet Res.* Canada; 2019; 21: e13260.

64. Gould MK, Huang BZ, Tammemagi MC, Kinar Y, Shiff R. Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data. *Am.*

J. Respir. Crit. Care Med. United States; 2021; 204: 445–453.

65. Rubin KH, Haastrup PF, Nicolaisen A, Möller S, Wehberg S, Rasmussen S, Balasubramaniam K, Søndergaard J, Jarbøl DE. Developing and Validating a Lung Cancer Risk Prediction Model: A Nationwide Population-Based Study. *Cancers (Basel).* Switzerland; 2023; 15.

66. Walter J, Kauffmann-Guerrero D, Muley T, Reck M, Fuge J, Günther A, Majeed RW, Savai R, Koch I, Dinkel J, others. Comparison of the sensitivity of different criteria to select lung cancer patients for screening in a cohort of German patients. *Cancer Med.* Wiley Online Library; 2023; 12: 8880–8896.

67. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, Hsieh LJ, Begg CB. Variations in lung cancer risk among smokers. *J. Natl. Cancer Inst.* United States; 2003; 95: 470–478.

68. Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. Development and validation of risk models to select ever-smokers for CT lung cancer screening. *Jama* American Medical Association; 2016; 315: 2300–2311.

69. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, Field J. The LLP risk model: an individual risk prediction model for lung cancer. *Br. J. Cancer* Nature Publishing Group; 2008; 98: 270–276.

70. Tammemägi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, Chaturvedi AK, Silvestri GA, Riley TL, Commins J, Berg CD. Selection criteria for lung-cancer screening. *N. Engl. J. Med.* United States; 2013; 368: 728–736.

71. Nguyen OTD, Fotopoulos I, Markaki M, Tsamardinos I, Lagani V, Røe OD. Improving Lung Cancer Screening Selection: The HUNT Lung Cancer Risk Model for Ever-Smokers Versus the NELSON and 2021 United States Preventive Services Task Force Criteria in the Cohort of Norway: A Population-Based Prospective Study. *JTO Clin. Res. Reports* 2024; 5: 100660.

72. Wood DE, Kazerooni EA, Baum SL, Eapen GA, Ettinger DS, Hou L, Jackman DM, Klippenstein D, Kumar R, Lackner RP, others. Lung cancer screening, version 3.2018, NCCN clinical practice guidelines in oncology. *J. Natl. Compr. Cancer Netw.* Harborside Press, LLC; 2018; 16: 412–441.

73. Chiu H-Y, Chao H-S, Chen Y-M. Application of Artificial Intelligence in Lung Cancer. *Cancers (Basel).* Switzerland; 2022; 14.

74. Gao Q, Yang L, Lu M, Jin R, Ye H, Ma T. The artificial intelligence and machine learning in lung cancer immunotherapy. *J. Hematol. \& Oncol.* Springer; 2023; 16: 55.

75. Capizzi G, Sciuto G Lo, Napoli C, Połap D, Woźniak M. Small lung nodules

detection based on fuzzy-logic and probabilistic neural network with bioinspired reinforcement learning. *IEEE Trans. Fuzzy Syst.* IEEE; 2019; 28: 1178–1189.

76. Hapke H, Nelson C. Building machine learning pipelines. O'Reilly Media; 2020.

77. Brownlee J. Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost [Internet]. 2021 [cited 2024 May 27].Available from: https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/.

78. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. *Proc. 22nd acm sigkdd Int. Conf. Knowl. Discov. data Min.* 2016. p. 785–794.

79. Geeks G for. Logistic Regression in Machine Learning [Internet]. 2023 [cited 2024 May 27].Available from: https://www.geeksforgeeks.org/understanding-logistic-regression/.

80. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning [Internet]. 2023Available from: https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.

81. Massachusetts Institute of Technology. Explained: Neural Networks [Internet]. 2017 [cited 2024 May 27].Available from: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414.

82. Schober P, Vetter TR. Logistic regression in medical research. *Anesth. Analg.* Wolters Kluwer Health; 2021; 132: 365.

83. Wüthrich M V, Merz M. Bayesian Methods, Regularization and Expectation-Maximization. *Stat. Found. Actuar. Learn. its Appl.* Springer; 2022. p. 207–266.

84. SHAP-documentation [Internet]. 2022.Available from: https://shap.readthedocs.io/en/latest/index.html.

85. SHAP documentation. 2022.

86. Heuvel T Ten. Opening the Black Box of Machine Learning Models: SHAP vs LIME for Model Explanation [Internet]. 2023 [cited 2024 May 28].Available from: https://medium.com/cmotions/opening-the-black-box-of-machine-learning-models-shap-vs-lime-for-model-explanation-d7bf545ce15f.

87. Bayes server. Asia [Internet]. [cited 2024 May 27].Available from: https://www.bayesserver.com/examples/networks/asia.

88. Matsumoto S, Carvalho RN, Ladeira M, Costa PCG, Santos LL, Silva D, Onishi M, Machado E, Cai K. UnBBayes: a java framework for probabilistic models in AI. *Java Acad. Res.* iConcept Press Annerley; 2011; : 34.

89. Norsys software corp. Example Bayes Net [Internet]. [cited 2024 May 29].Available from: https://www.norsys.com/WebHelp/NETICA/X_Example_Bayes_Net.htm.

90. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B* Wiley Online Library; 1988; 50: 157–194.

91. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. Springer International Publishing; 2019.Available from: https://books.google.dk/books?id=d2WCwgEACAAJ.

92. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N. Engl. J. Med.* Mass Medical Soc; 1980; 302: 1109–1117.

93. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic Progn. Res.* BioMed Central; 2019; 3: 1–8.

94. Horeweg N, van der Aalst CM, Vliegenthart R, Zhao Y, Xie X, Scholten ET, Mali W, Thunnissen E, Weenink C, Groen HJM, others. Volumetric computed tomography screening for lung cancer: three rounds of the NELSON trial. *Eur. Respir. J.* Eur Respiratory Soc; 2013; 42: 1659–1667.

95. Lebrett MB, Balata H, Evison M, Colligan D, Duerden R, Elton P, Greaves M, Howells J, Irion K, Karunaratne D, others. Analysis of lung cancer risk model (PLCOM2012 and LLPv2) performance in a community-based lung cancer screening programme. *Thorax* BMJ Publishing Group Ltd; 2020; 75: 661–668.

96. Tammemägi MC, Church TR, Hocking WG, Silvestri GA, Kvale PA, Riley TL, Commins J, Berg CD. Evaluation of the lung cancer risks at which to screen ever- and never-smokers: screening rules applied to the PLCO and NLST cohorts. *PLoS Med.* 2014; 11: e1001764.

97. Tammemägi MC, Darling GE, Schmidt H, Llovet D, Buchanan DN, Leung Y, Miller B, Rabeneck L. Selection of individuals for lung cancer screening based on risk prediction model performance and economic factors - The Ontario experience. *Lung Cancer* Ireland; 2021; 156: 31–40.

98. European Union. Attitudes towards the impact of digitisation and automation on daily life [Internet]. 2017 [cited 2024 Sep 12].Available from: https://europa.eu/eurobarometer/surveys/detail/2160.

99. Center PR. How americans think about artificial intelligence [Internet]. 2022 [cited 2024 Sep 12].Available from: https://www.pewresearch.org/internet/2022/03/17/how-americans-think-about-artificial-intelligence/.

100. Pew Research center. 60% of americans would be uncomfortable with provider relying on AI in their own health care [Internet]. 2022Available from: https://www.pewresearch.org/wp-content/uploads/sites/20/2023/02/PS_2023.02.22_AI-health_REPORT.pdf.

101. Park HJ. Patient perspectives on informed consent for medical AI: A web-based experiment. *Digit. Heal.* SAGE Publications Sage UK: London, England; 2024; 10: 20552076241247936.

102. Pruski M. AI-Enhanced Healthcare: Not a new Paradigm for Informed Consent. *J. Bioeth. Inq.* Springer; 2024; : 1–15.

103. Pinsky P. Electronic Health Records and Machine Learning for Early Detection of Lung Cancer and Other Conditions: Thinking about the Path Ahead. Am. J. Respir. Crit. Care Med. United States; 2021. p. 389–390.

104. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation." *AI Mag.* 2017; 38: 50–57.

105. Ali N, Lifford KJ, Carter B, McRonald F, Yadegarfar G, Baldwin DR, Weller D, Hansell DM, Duffy SW, Field JK, others. Barriers to uptake among high-risk individuals declining participation in lung cancer screening: a mixed methods analysis of the UK Lung Cancer Screening (UKLS) trial. *BMJ Open* British Medical Journal Publishing Group; 2015; 5: e008254.

106. Menakuru SR, Dhillon VS, Beirat AF, Hanna NH. Patient perception and adherence to lung cancer screening in those who meet eligibility criteria. American Society of Clinical Oncology; 2023.

107. E boks. E boks: Denmark's largest and most popular postbox [Internet]. [cited 2024 May 28].Available from: https://private.e-boks.com/danmark/en/.

108. Brown L, Agrawal U, Sullivan F. Using electronic medical records to identify potentially eligible study subjects for lung cancer screening with biomarkers. *Cancers (Basel).* MDPI; 2021; 13: 5449.

109. Crosbie PA, Balata H, Evison M, Atack M, Bayliss-Brideaux V, Colligan D, Duerden R, Eaglesfield J, Edwards T, Elton P, others. Implementing lung cancer screening: baseline results from a community-based 'Lung Health Check'pilot in deprived areas of Manchester. *Thorax* BMJ Publishing Group Ltd; 2019; 74: 405–409.

110. Criner GJ, Agusti A, Borghaei H, Friedberg J, Martinez FJ, Miyamoto C, Vogelmeier CF, Celli BR. Chronic Obstructive Pulmonary Disease and Lung Cancer: A Review for Clinicians. *Chronic Obstr. Pulm. Dis. (Miami, Fla.)* United States; 2022; 9: 454–476.

111. Lowry KP, Gazelle GS, Gilmore ME, Johanson C, Munshi V, Choi SE, Tramontano AC, Kong CY, McMahon PM. Personalizing annual lung cancer screening for patients with chronic obstructive pulmonary disease: A decision analysis. *Cancer* United States; 2015; 121: 1556–1562.

112. The Danish Health Data Authority. SKS-browser [Internet]. 2023 [cited 2023 Dec 13].Available from: https://medinfo.dk/sks/brows.php.

113. Jakobsen E, Rasmussen TR. The Danish Lung Cancer Registry. *Clin. Epidemiol.* New Zealand; 2016; 8: 537–541.

114. Mooney G. The Danish health care system: it ain't broke… so don't fix it. *Health Policy (New. York).* Elsevier; 2002; 59: 161–171.

115. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* Springer; 1992; 9: 309–347.

116. Dempster A, Laird N, Rubin D, Dempster AP, Laird NM, Rubin D, others. Likelihood from incomplete data via the em algorithm. *JR Stat. Soc. B. v39 i1* 1977; : 1–38.

117. Cobb BR, Rumí R, Salmerón A. Bayesian Network Models with Discrete and Continuous Variables. In: Lucas P, Gámez JA, Salmerón A, editors. *Adv. Probabilistic Graph. Model.* [Internet] Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 81–102Available from: https://doi.org/10.1007/978-3-540-68996-6_4.

118. Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes for classification learning. *Ijcai* 1993. p. 1022–1029.

119. Arias M, Pérez-Mart\'\in J, Luque M, D\'\iez FJ. OpenMarkov, an Open-Source Tool for Probabilistic Graphical Models. *IJCAI* 2019. p. 6485–6487.

120. Danish Clinical Quality Program. DrCOPD [Internet]. 2023 [cited 2023 Apr 24].Available from: https://www.rkkp.dk/kvalitetsdatabaser/databaser/dansk-register-for-kronisk-obstruktiv-lungesygdom/.

121. Tammemägi MC. Selecting lung cancer screenees using risk prediction models-where do we go from here. *Transl. lung cancer Res.* China; 2018; 7: 243–253.

122. Choi E, Ding VY, Luo SJ, Ten Haaf K, Wu JT, Aredo J V, Wilkens LR, Freedman ND, Backhus LM, Leung AN, others. Risk model--based lung cancer screening and racial and ethnic disparities in the US. *JAMA Oncol.* American Medical Association; 2023; 9: 1640–1648.

123. Tourangeau R, Yan T. Sensitive questions in surveys. *Psychol. Bull.* American Psychological Association; 2007; 133: 859.

124. Kaur G, Begum R, Thota S, Batra S. A systematic review of smoking-related epigenetic alterations. *Arch. Toxicol.* Springer; 2019; 93: 2715–2740.

125. Bahado-Singh R, Vlachos KT, Aydas B, Gordevicius J, Radhakrishna U, Vishweswaraiah S. Precision Oncology: Artificial Intelligence and DNA Methylation Analysis of Circulating Cell-Free DNA for Lung Cancer Detection. *Front. Oncol.* Switzerland; 2022; 12: 790645.

126. Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, Caughey AB, Davis EM, Donahue KE, Doubeni CA, Kubik M, others. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *Jama* American Medical Association; 2021; 325: 962–970.

127. van der Aalst CM, Ten Haaf K, de Koning HJ. Implementation of lung cancer screening: what are the main issues? *Transl. Lung Cancer Res.* AME Publications; 2021; 10: 1050.

128. Vonder M, der Aalst C, Hubert J, Moldovanu D, Schmitz A, Delorme S, Gratama JW, Silva M, de Koning H, Oudkerk M. MA19. 06 Artificial Intelligence as Concurrent Reader in Prospective European Lung Cancer Screening (4-IN-THE-LUNG-RUN) Trial. *J. Thorac. Oncol.* Elsevier; 2023; 18: S172.

129. Conjeti S. Transforming Healthcare: A Step-by-Step Guide to Building and Deploying AI Medical Devices [Internet]. 2023.Available from: https://www.linkedin.com/pulse/transforming-healthcare-step-by-step-guide-building-ai-conjeti/.

130. Craddock M, Crockett C, Mcwilliam A, Price G, Sperrin M, Van Der Veer SN, Faivre-Finn C. Evaluation of prognostic and predictive models in the oncology clinic. *Clin. Oncol.* Elsevier; 2022; 34: 102–113.

131. de Sande D, Van Genderen ME, Smit JM, Huiskens J, Visser JJ, Veen RER, van Unen E, Hilgers O, Gommers D, van Bommel J. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Heal. \& care informatics* BMJ Publishing Group; 2022; 29.