Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

A Validated Clinical Risk Prediction Model for Lung Cancer in Smokers of All Ages and Exposure Types: A HUNT Study

Maria Markaki [a], Ioannis Tsamardinos [a,b], Arnulf Langhammer [c], Vincenzo Lagani [a,b], Kristian Hveem [c,g], Oluf Dimitri Røe [d,e,f,*]

[a] University of Crete, Department of Computer Science, Voutes Campus, Heraklion, GR 70013, Greece
[b] Gnosis Data Analysis PC, Palaiokapa 64, Heraklion, GR 71305, Greece
[c] HUNT Research Centre, Department of Public Health and Nursing, Norwegian University of Science and Technology, Forskningsvegen 2, Levanger, NO 7600, Norway
[d] Norwegian University of Science and Technology, Department of Clinical Research and Molecular Medicine, Prinsesse Kristinsgt. 1, Trondheim, NO 7491, Norway
[e] Levanger Hospital, Nord-Trøndelag Hospital Trust, Cancer Clinic, Kirkegata 2, Levanger, NO 7600, Norway
[f] Clinical Cancer Research Center, Department of Clinical Medicine, Hobrovej 18-22, Aalborg, DK 9000, Denmark
[g] K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health an Nursing, Norwegian University of Science and Technology, NO 7491 Trondheim, Norway

ARTICLE INFO

Article history:
Received 25 January 2018
Received in revised form 19 March 2018
Accepted 22 March 2018
Available online 30 March 2018

Keywords:
Early diagnosis
Lung cancer prediction
Ever-smokers
All smokers
All ages
Data-driven
Feature selection
External validation

ABSTRACT

Lung cancer causes >1·6 million deaths annually, with early diagnosis being paramount to effective treatment. Here we present a validated risk assessment model for lung cancer screening.

The prospective HUNT2 population study in Norway examined 65,237 people aged >20 years in 1995–97. After a median of 15·2 years, 583 lung cancer cases had been diagnosed; 552 (94·7%) ever-smokers and 31 (5·3%) never-smokers. We performed multivariable analyses of 36 candidate risk predictors, using multiple imputation of missing data and backwards feature selection with Cox regression. The resulting model was validated in an independent Norwegian prospective dataset of 45,341 ever-smokers, in which 675 lung cancers had been diagnosed after a median follow-up of 11·6 years.

Our final HUNT Lung Cancer Model included age, pack-years, smoking intensity, years since smoking cessation, body mass index, daily cough, and hours of daily indoors exposure to smoke. External validation showed a 0·879 concordance index (95% CI [0·866–0·891]) with an area under the curve of 0·87 (95% CI [0·85–0·89]) within 6 years. Only 22% of ever-smokers would need screening to identify 81·85% of all lung cancers within 6 years.

Our model of seven variables is simple, accurate, and useful for screening selection.

1. Introduction

Lung cancer (LC) is the leading cause of cancer mortality worldwide (Torre et al., 2016), and early diagnosis is paramount for increasing survival. The National Lung Screening Trial (NLST) showed that low-dose high-resolution computed axial tomography (CT) scanning of heavy smokers (>30 pack-years, <15 years quit time) aged 55–74 at inclusion time and at 6 years of follow-up reduced LC mortality by 20% (National Lung Screening Trial Research Team, 2011). However, these simple criteria are relatively ineffective. First, only an estimated 26·7% of those who develop LC in a general population cohort fulfil the NLST inclusion criteria for CT screening (Pinsky and Berg, 2012). Second, out of those included, false-positive or indolent LCs counted for 96·4% and 18% of cases, respectively. In addition, the potential danger of unnecessary invasive and potentially dangerous procedures, the psychological burden of a false-positive finding, and risks associated with CT screening are not negligible (Patz et al., 2014; Rampinelli et al., 2017). Specifically, the risk for LC induced by the radiation from the CT screening is estimated to be between 24 and 81/100000 cases after 10 years of CT screening (Rampinelli et al., 2017).

The above arguments suggest a pressing need for improving the NLST criteria for effective CT screening. In a European Union position statement recently published in *Lancet Oncology*, risk stratification is one of the keys to ensure the successful implementation of future low-dose CT screening programmes in Europe (Oudkerk et al., 2017).

Several multivariable risk prediction models have been proposed to improve the selection of people for LC screening (Ten Haaf et al., 2017; Tammemagi et al., 2013). In addition to NLST's pack-years, quit-time, and age, they consider other risk factors such as history of respiratory diseases, exposure to occupational dust (asbestos, coal, silica), socioeconomic status, body mass index (BMI), history of cancer, race, education,

[*] Corresponding author at: Norwegian University of Science and Technology, Department of Clinical Research and Molecular Medicine, Prinsesse Kristinsgt. 1, Trondheim, NO 7491, Norway.
E-mail address: oluf.roe@ntnu.no. (O.D. Røe).

https://doi.org/10.1016/j.ebiom.2018.03.027

forced expiratory volume and biochemical parameters such as carcinoembryonic antigen, alpha-fetoprotein, and C-reactive protein (Katki et al., 2016; Wu et al., 2016; Muller et al., 2017).

However, these models and corresponding studies also have a variety of potential issues such as age cutoffs, inclusion of mainly heavy smokers, restricted and/or empirical inclusion of predictors and listwise exclusion of cases with missing data, all of which call into question the transferability of these models to clinical practice. Can we create a model that can reliably predict LC across ages and smoking burdens?

To address this challenge, we developed a novel LC risk prediction model for CT screening based on data from a large, prospective, population-based study in Norway of 65,237 people aged 20–100 with a median follow-up time of 15·2 years. Multivariable statistical methods identified a minimal set of required factors to achieve optimal prediction. The model has been successfully externally validated on a larger independent cohort. Our study furthers the state-of-the-art by developing a model trained on a population with a wider age group, which includes light smokers, has a relatively long follow-up median time, performs data-driven selection of predictors, and does not exclude cases with missing data (handled with multiple imputation).

2. Methods

2.1. Ethics

Participants included in HUNT2 and Cohort of Norway (CONOR) all gave their written consent. The Norwegian Data Inspectorate and the Regional Committees for Medical Research Ethics approved each individual study.

2.2. Discovery Dataset: The HUNT2 Population

The Nord-Trøndelag Health Study (HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health.

From 1995 to 1997, HUNT2 invited 93,898 residents of Nord-Trøndelag County in Norway, aged 20 years or more, to participate in a health survey, and ≈70% ($n = 65{,}237$) responded (Krokstad et al., 2013). The data were collected through questionnaires on demographic characteristics, medical history, and lifestyle (199 clinical variables/questions selected from the HUNT2 Baseline Questionnaires 1 and 2 and Measurements NT2BLQ1, NT2BLQ2, and NT2BLM, respectively) (HUNT, 2018). In 2012, our group was granted access to analyse the HUNT2 data to identify LC cases and establish the HUNT2 discovery dataset. We also linked the national 11-digit personal identification number of each participant to the Norwegian Cancer and Death Cause Registry. The diagnosis code of the International Classification of Diseases 7162·1 was used. Individuals who died or migrated were modelled as censored at the time they left the study; the latest follow-up day for all participants was December 31, 2011. Those who developed other cancer types during the follow-up period ($n = 6821$), had a LC diagnosis before their participation in HUNT2 ($n = 16$), or did not answer any questionnaire in HUNT2 ($n = 57$) were excluded from the current study, resulting in a subset of 58,343 eligible participants.

2.3. Variables

We identified 36 potential predictors out of the 199 including age, sex, education, BMI, history of previous cancer, asthma, heart attack, stroke, fractures, self-perceived health, various heart- and lung-related symptoms, anxiety, muscle pain, detailed smoking history (including indoor smoke exposure in hours and smoke exposure as a child), asthma medication, daily coffee use, and physical activity (Table 1, Supplementary Table S1a). The selection criteria were based on known risk factors for LC as well as factors associated with other smoke-related diseases. Ever-smokers were defined as those who responded positively to the question, “Smoke daily now or ever?”; those who answered negative were defined as never-smokers.

2.4. Validation Dataset: The CONOR Population

The risk prediction model learned from the HUNT2 analysis was applied and externally validated on the ever-smokers in the CONOR database. CONOR constitutes a national database of ten regional prospective population-based studies of 173,236 individuals aged >19 that use the same questionnaires as HUNT2 (Naess et al., 2008). Urban population from the largest cities of Norway as well as rural population is represented. It also includes some participants born in non-European countries (HUBRO Study) (Sogaard et al., 2004) representing 1·4% of ever-smokers and an unknown fraction of indigenous Sami people in studies from northern Norway (Naess et al., 2008).

All participants with complete predictor data in CONOR were included while all HUNT2 participants ($n = 65{,}018$) and never-smokers ($n = 21{,}649$) were excluded. To simulate a true screening setting, those with previous history or subsequently diagnosed with other cancers were not excluded.

2.5. Statistical Analysis

The original variables were non-linearly transformed whenever necessary; specifically age, pack-years, quit-time, BMI, and hours of indoors exposure to smoke. Missing values were imputed using multiple imputation with predictive mean matching (R package mice) (van Buuren and mice, 2011), resulting in 30 complete datasets. For each of them, 200 bootstrap datasets were generated, and backwards feature selection with the Akaike Information Criterion was performed on every set using the R package rms (Harrell, 2001). Second-order interaction terms were also tested for inclusion. The predictors that were returned by the above procedure in the majority of datasets were selected in the final model, and their regression coefficients were calculated according to “Rubin's Rules” (Heymans et al., 2007). Internal validation was performed with the bootstrap method; for all metrics, median and interquartile range in the multiple imputed datasets are reported (robust methods) (Marshall et al., 2009). Discriminative ability was measured by the concordance index (C-index) metric. Calibration, i.e. agreement between predicted and observed risks across subgroups of the population, was evaluated by the predictiveness curve (Pepe et al., 2008). An online risk calculator was created; the electronic version of the calculator is available at (http://mensxmachina.org/en/HUNT-NTNU-lung-cancer-risk-calculator/). The results of the modelling process are also presented by a nomogram where the relative importance of each predictor is depicted and where the 5-, 10-, or 15-year estimates of the LC risk could be calculated. The statistical methodology is presented in detail in the Supplementary Appendix (Supplementary Figs. S1–4 and Table S1a–b). The analysis conforms to the reporting standards of STARD/TRIPOD (Moons et al., 2015; Bossuyt et al., 2015), and is depicted in Fig. 1.

2.6. Role of the Funding Source

The funding sources had no role in study conception, design, interpretation of the data, writing of the report, or decision to submit the paper for publication. The corresponding author confirms that he had full access to all data in the study and final responsibility for the decision to submit for publication.

3. Results

In the HUNT2 discovery cohort ($n = 58{,}343$; 800,845 person-years), 57·5% of individuals were ever-smokers ($n = 33{,}521$; 469,404 person-years), and 583 were diagnosed with LC during follow-up (median

Table 1
All 36 variables included in the backwards feature selection analysis. In univariate analysis, all variables except those indicated in green had significant *p* values regarding the risk for LC diagnosis ($p < 0.05$, see Supplementary Table S1b). However, in multivariate analysis, only the red ones were selected and included in the final model.

Basic features	History of previous disease	Symptoms	Smoking history	Medication	Physical activity
🟥 Age (PartAg)	🟦 Cancer ever (CaEv)	🟥 Cough daily (CougDy)	🟥 Pack years (SmoPackYrs)	🟦 Medication daily last year (MedDyLy)	🟦 Vigorous exercise per week (ExeHarDuLy)
🟥 Body mass index (BMI)	🟦 Asthma ever (AstEv)	🟦 Cough phlegm (CougPhle)	🟥 Smoke intensity (SmoCigDyN)	🟦 Asthma medication ever (AsthMedEV)	🟩 Light exercise per week (ExeLigDuLy)
🟦 Sex	🟦 Heart attack ever (CarInfEv)	🟦 Dyspnea last year (DysLy)	🟥 Indoor smoke exposure hours (SmoExpH)	🟦 Cups of brewed coffee daily (DriCofBoilNDy)	
🟦 Education (Educ)	🟦 Self-perceived health (Healt)	🟦 Wheezing last year (WheeDysLy)	🟥 Quit time Years (SmoDyCesDu)		
	🟩 Stroke (ApoplEv)	🟦 Angina pectoris (CarAngEv)	🟦 Cigar/cigarillos daily (SmoCigarDy)		
	🟩 Fracture hip ever (FracHipEv)	🟦 Palpitations last year (CarTachLY)	🟦 Current smoker (SmoCigDy)		
	🟩 Fracture wrist ever (FracWriEv)	🟦 Muscle pain last year (MSPaLY)	🟦 Never smoker (SmoDyNev)		
		🟦 Muscle pain duration in years (MSPaDuYrs)	🟦 Age start smoking (SmoDyAg)		
		🟦 Insomnia frequency (InsomF)	🟦 Smoking duration (SmoDyDuEd)		
		🟩 Anxiety-depression (HADSTotExtr)	🟩 Smoke exposure as child (SmoExpCh)		

*The variables were selected from the HUNT2 Baseline Questionnaire 1, 2 and Measurements (NT2BLQ1, NT2BLQ2, and NT2BLM respectively, see link https://www.ntnu.no/hunt/variabler and Supplementary Table S1).

follow-up, 15·2 years), corresponding to a 16-year cumulative incidence of ~1%. Among LC cases, 552 (94·7%) were ever-smokers and 31 (5·3%) never-smokers; thus, incidence rates of LC per 10,000 person-years in never- and ever-smokers were 0·7 and 16·5, respectively (Supplementary Table S1b).

3.1. Univariate Analysis

In univariate analysis, 30 of the 36 candidate variables indicated a higher LC risk; among these were higher age, male gender, individuals with lower education, lower BMI, individuals who had smoked more pack-years, more cigarettes per day, and for longer durations. However, several unexpected variables such as any incidence of heart attack, current self-perceived health, muscle pain, insomnia, and daily intake of cups of brewed coffee were also statistically significant (Table 1 and Supplementary Table S1b).

3.2. Multivariable Risk Prediction Model (HUNT Lung Cancer Model)

Seven variables were selected by the backward selection procedure: age, total smoking burden (pack-years), quit time ("If you previously smoked, how long has it been since you stopped? Number of years"), daily cough ("Do you cough daily during periods of the year?Yes or No"), BMI, hours of indoor smoke exposure ("How long are you usually in a smoky room each day? Number of hours"), and smoking intensity (number of cigarettes per day) (Table 2, Supplementary Table S1a). Hazard ratios for LC increased with age, pack-years, hours of daily exposure to smoke, and cough daily during periods of the year. Hazard ratios decreased with increasing BMI, quit time, and smoking intensity given a fixed number of pack-years (e.g. smoking 20 pack-years by 40 cigarettes per day for 10 years is less deleterious than 10 cigarettes per day for 40 years) (Table 2, Supplementary Table S4).

Using this model, the median C-index was 0·903 in the full cohort (including never-smokers), whereas for ever-smokers in the same group, the median C-index was 0·869 (Table 3, Tables S4 and S12). Calibration was satisfactory, with observed risks very close to the predicted (Supplementary Figs. S3 and S4). However, it was not possible to develop a model for the never-smokers with predictions better than random guessing (data not shown), so the final model, the "HUNT Lung Cancer Model", included only ever-smokers. The model is presented as a nomogram (Fig. 2a). An online risk calculator were also created; the electronic version of the calculator is available at (http://mensxmachina.org/en/HUNT-NTNU-lung-cancer-risk-calculator/).

The risk predictions were stratified into low-, medium-, and high-risk groups using the 50% and 84% quantiles, as previously proposed (Royston and Altman, 2013), and visualized by Kaplan–Meier curves, showing statistically significant differences between the groups by a log-rank test (Fig. 2b, p = 0·0008). To estimate the absolute risk per individual, the baseline risk of LC diagnosis (the survival function) was estimated according to van Houwelingen (Supplementary Appendix). Sex did not contribute significantly but was included for adjustment (Table 3; see comment in Discussion). The detailed analysis of two separate gender-specific models is described in the Supplementary Tables S7 and S8.

3.3. External Validation

In the external validation dataset (CONOR), among participants with complete data ($n = 67{,}036$; 807,701 person-years), 67·7% were ever-

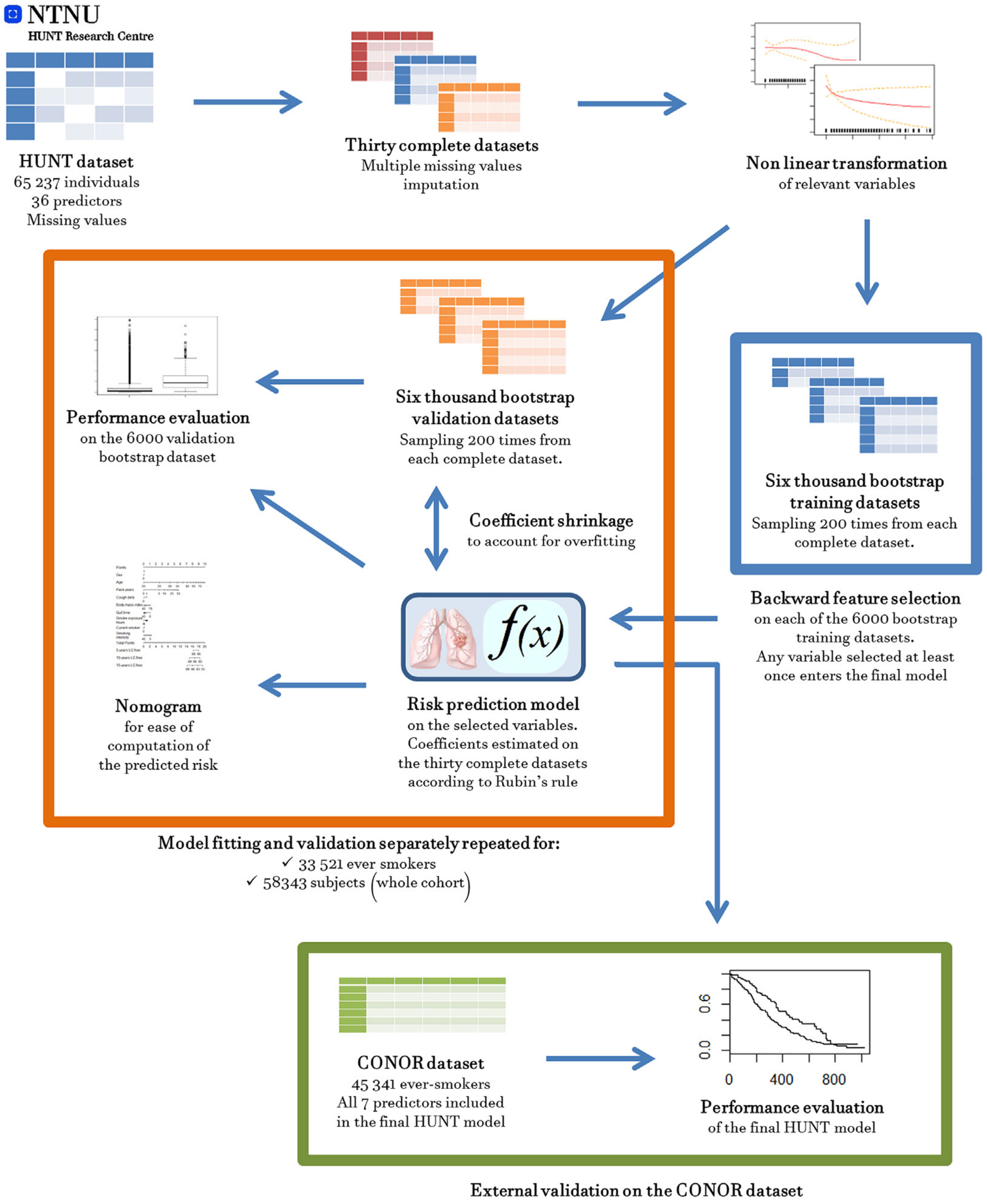
NTNU
HUNT Research Centre
HUNT dataset
65 237 individuals
36 predictors
Missing values
Thirty complete datasets
Multiple missing values imputation
Non linear transformation
of relevant variables
Performance evaluation
on the 6000 validation bootstrap dataset
Six thousand bootstrap validation datasets
Sampling 200 times from each complete dataset.
Coefficient shrinkage
to account for overfitting
Six thousand bootstrap training datasets
Sampling 200 times from each complete dataset.
Backward feature selection
on each of the 6000 bootstrap training datasets.
Any variable selected at least once enters the final model
Nomogram
for ease of computation of the predicted risk
$f(x)$
Risk prediction model
on the selected variables.
Coefficients estimated on the thirty complete datasets according to Rubin's rule
Model fitting and validation separately repeated for:
✓ 33 521 ever smokers
✓ 58343 subjects (whole cohort)
CONOR dataset
45 341 ever-smokers
All 7 predictors included in the final HUNT model
0.6
0.0
0
400
800
Performance evaluation
of the final HUNT model
External validation on the CONOR dataset

Fig. 1. Schematic representation of the analysis protocol. The analyses were performed on 36 selected predictors measured on a cohort of 58,343 participants involved in the HUNT study. Thirty distinct datasets were created after missing data imputation, and selected variables were transformed in a non-linear fashion. For each of the 30 complete datasets, 200 bootstrapped datasets were created, leading to a total number of 6000 training datasets. A backward feature selection (with Akaike Information Criterion as a stopping rule) was repeated on each training dataset for selecting the most relevant risk factors for lung cancer. A final model was built over all variables chosen at least once during the feature selection procedure, and by using Rubin's rule for estimating model coefficients from the 30 complete datasets. A total of 6000 separate bootstrapped validation datasets were created for assessing the predictiveness of the model and for correcting for overfit. Both the fitting of the model and its evaluation were repeated for the whole cohort and for ever-smokers only.

smokers (n = 45,341, 566,579 person-years) (Supplementary Table S10). Median pack-years were 11·5. In this cohort, there were 709 LC events until 2011, 675 (95·5%) and 34 (4·5%) among ever- and never-smokers, respectively, with a median follow-up-time of 11·6 years (Supplementary Tables S8–S9). The incidence rates of LC per 10,000 person-years in never- and ever-smokers were 1·4 and 11·9, respectively.

Applying the HUNT Lung Cancer Model in the ever-smokers of the validation cohort resulted in a C-index 0·879 (95% confidence interval [CI] 0·866–0·891) (Table S11), close to the one estimated on the HUNT data. Calibration was satisfactory (Supplementary Figs. S3 and S4). The area under the curve (AUC) within 6 years was 0·87 (95% CI 0·85–0·89). Setting the risk threshold to correspond to the 16% quantile of the risk of events in HUNT (medium and high risk) equalled a risk of 1·75% for developing LC within 16 years (>15 points in the nomogram) and a risk of 0·64% for developing LC within 6 years (Fig. 2b). Based on this threshold, 221 of 270 LC events within 6 years (sensitivity and specificity of 81.85% and 78.31%, respectively) and 527 of 675 events within ~20 years (sensitivity and specificity of 78.07% and 78.82%, respectively) were correctly predicted. More specifically, using this threshold, one would need to examine 10,000 (22%) out of 45,387 ever-smokers to identify 81·85% of future LC events in a 6-year period or 78% of future events in a 20-year period (median 11·6 years) (Fig. 4d, Table 4).

Table 2
The HUNT Lung Cancer Model. Variables and questions to participants. Cox prediction model of lung cancer risk for 33,521 HUNT2 participants who had ever smoked[*] and did not develop any other type of cancer in a mean follow-up time of 13·2 years.

Variable	Questions to participants	Hazard ratio (95% CI)	P value	Beta coefficient
Sex	-	1·128 (0·941–1·352)	0·188	0·1205819
Age[a]	Age at participation at screening	0·135 (0·098–0·186)	<0·001	-2·0020557
Pack-years (log)	Estimated number of pack-years	3·200 (2·451–4·176)	<0·001	1·1630181
Smoking quit time, years (log)	If you previously smoked, how long has it been since you stopped? (Number of years)	0·786 (0·705–0·876)	<0·001	-0·2407998
Body mass index (log)	BMI	0·288 (0·153–0·539)	<0·001	-1·2462656
Cough daily, yes vs no	Do you cough daily during periods of the year?	1·501 (1·250–1·802)	<0·001	0·4059355
Smoke exposure, hours (log)	How long are you usually in a smoky room each day? (Number of hours)	1·181 (1·062–1·313)	0·002	0·1663201
Smoking intensity per 1 cigarette increase	How many cigarettes do you or did you usually smoke daily?	0·971 (0·951–0·991)	0·004	-0·0295406

[*] To calculate the 16-year lung cancer risk in one person with the use of categorical variables, multiply the beta coefficient of the variable by 1 if the factor is present and by 0 if it is absent. For continuous variables other than age, multiply their value – or their log value if indicated – by the beta coefficient of the variable. For age, calculate its contribution by dividing by 100, exponentiated by the power −1, and multiply by the beta coefficient of the variable. Calculate the sum of all previously calculated beta coefficient products; this sum is represented as $X\beta$. To obtain the person's 16-year LC risk, calculate $1 - 0.06^{\exp(X\beta)}$. CI denotes confidence interval.

[a] Age had a non-linear association with LC and was transformed as (100/Age).

3.4. HUNT Lung Cancer Model and NLST

The number of people in CONOR (validation population) fulfilling NLST criteria was 2081; only 69 (25·6%) out of the 270 that developed LC within 6 years were included in this group out of 2081 people. We contrasted the performance of the HUNT Lung Cancer Model with the NLST criteria, given the same number of individuals screened ($n = 2081$), by selecting the top 2081 highest-risk individuals in CONOR, as assessed by the model (Table 5). The proposed model's criteria identified 103 vs 69 of NLST out of 270 cases showing an improved sensitivity (38.14% vs 25.6%, $P = 0.0216$) and positive predictive value (4.95% vs 3.3%, $P < 0.000001$), with the same specificity (95.61% vs 95.5%, $P = 0.7321$) and similar negative predictive value (99.6% vs. 99.5%, $P = 0.95374$).

3.5. How to Apply the Model

A user-friendly LC risk calculator for 6 and 16 years was created which has good prediction accuracy and seems well calibrated both in HUNT and in CONOR (Supplementary Fig. S3). The nomogram provided (Fig. 2a) can also be used to calculate LC risk. The nomogram visually depicts the relationship between model variables and the LC risk: the length of each variable axis is proportional to the contribution of this variable to the total risk. One simply adds the contributions of all the variables in the nomogram and the result reflects the personal risk of this individual. In our calculations we set the risk threshold to correspond to the 16% quantile of the risk of events in HUNT (medium and high risk, Fig. 2b) equalling a risk of at least 1·75% for developing LC within 16 years and at least 0.64% in 6 years (>15 points in the

Table 3
Key differences of the HUNT Lung Cancer Model over externally validated risk prediction models developed in prospective population-based cohorts. AUC refers to prediction of 1-, 5- (EPIC), or 6-year cancer risk (PLCO, HUNT2).

Key studies	LLPi	EPIC	PLCO$_{M2012}$	HUNT2	CONOR
Reference	Marcus et al., 2015, Raji et al., 2012	Hoggart et al., 2012	Weber et al., 2017, Tammemagi et al., 2013	Discovery cohort	Validation Cohort
Study group characteristics					
Cohort type	Random selection (n=8760)	Multi-country health study (n=399 393)	Multicentre randomized screening (n=80 375)	**One county 70% of total adult population (n=65 240)**	**One country, 11 health studies ever-smokers (n=45 341)**
Age limit	45–79	35+	55–74	**= 20**	**= 20**
Median Pack-years	18·9	≈30[a]	≈30	**10·3**	**11·5**
Never-smokers analysed	Yes	Yes	No	**Yes (n=24 725)**	**Not applicable**
Follow-up, years	8·7 mean	5 max	6 max	**13·2 mean**	**16 max**
Feature selection	Yes backward	Yes, based on AUC and tdNRI	No, pre-specified	**Yes backward**	**Not applicable**
Number of variables	14 (6 selected)	12 (4 selected)	11	**36 (7 selected)**	**7**
Coding of non-linearities of continuous variables	No	Yes, including stratification	Yes	**Yes**	**Yes**
Report on missing data	Yes	Yes	No	**Yes**	**Yes**
MI[c]	No	No	No	**Yes**	**Not applicable**
MI with feature selection[c]	No	No	No	**Yes**	**Not applicable**
Internal validation	Yes	No	Yes	**Yes**[b]	**Not applicable**
External validation	Yes	EPIC test set	Yes	**Yes**	
Discriminatory power (AUC[d] and/or C-index[e])					
	C-index	**AUC**	**AUC**	**C-index / AUC**	
Total Population	0·849	NR	NR	**0·903**	
Ever-smokers	NR **AUC 5y)**	NR	0·803 (6 y)	**0·869**	
External validation	0·67, 0·76, 0·82	0·787 (5 y) (Vlaanderen et al., 2014)	0·797 (6 y)	**0·879 / 0·87 (6 y)**	

NR = not reported.

[a] Years of smoking more than >15 cigarettes per day.

[b] Bootstrap in each of 30 multiply imputed datasets.

[c] MI = multiple imputation.

[d] Area under the receiver operating curve.

[e] Concordance index.

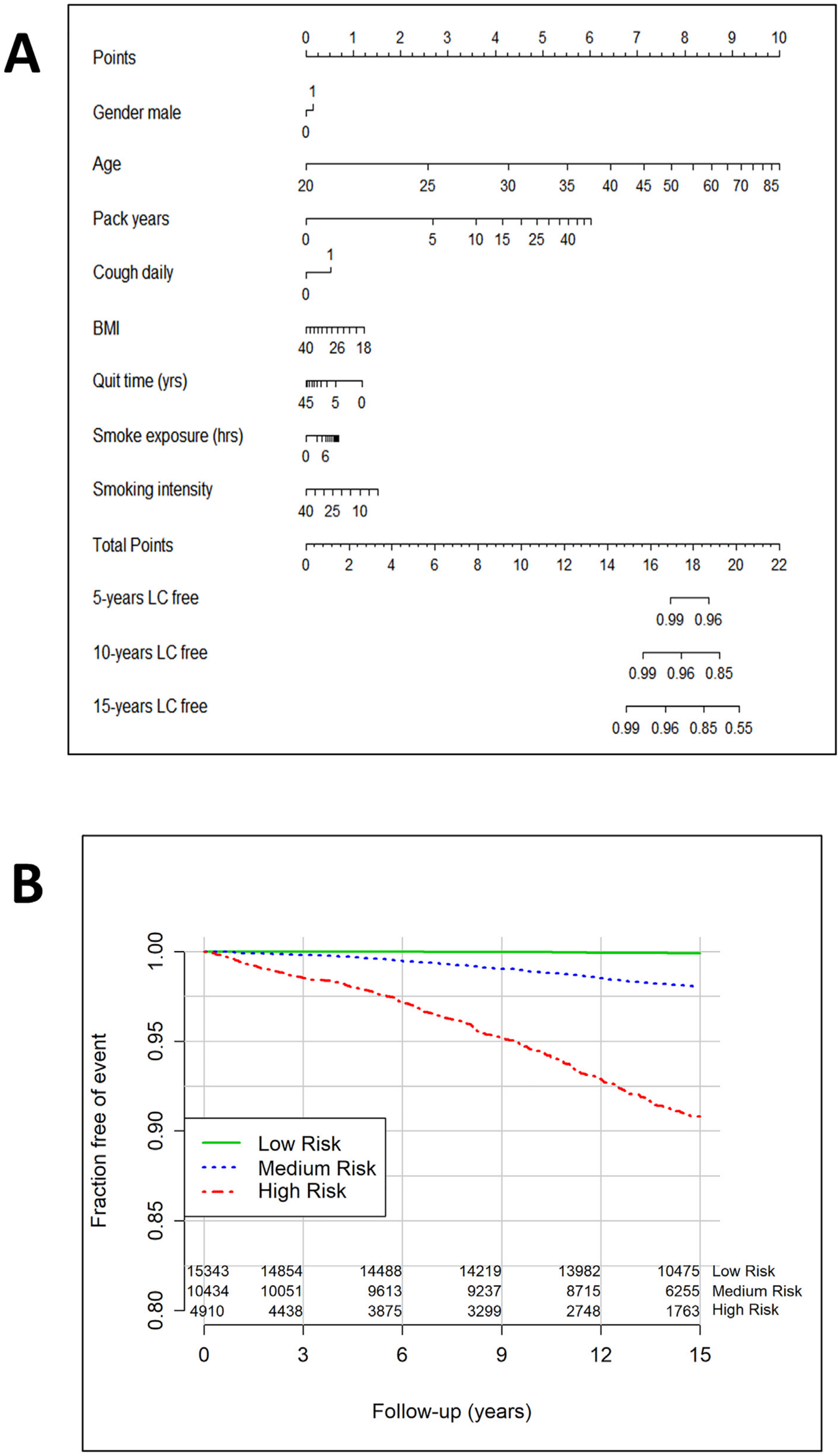

Fig. 2. A. Nomogram to calculate the calculate the personal 5-, 10-, and 15-year risk of lung cancer risk with the use of seven independent factors discovered by backwards feature selection. B. Low-, medium-, and high-risk groups for lung cancer according to the risk prediction model for ever-smokers. The Kaplan–Meier curves are plotted for risk groups defined from 50% and 84% quantiles. Differences among the three curves were highly significant ($p = 0{\cdot}0008$) according to the log-rank test. The number of event-free participants in every risk group at different time points is shown above the x-axis.

nomogram, Fig. 2a). By applying this threshold, using either the HUNT Lung Cancer Model nomogram or the online calculator, a 40-year old person with 15 pack-years and full-score contribution with low BMI (e.g. 22), low smoke intensity (e.g. 10 cigarettes per day), periodical or daily cough, and many hours of indoor smoke exposure (e.g. 10 h, total risk score > 15, LC risk = 1·77% at 16 years and 1.39% in 6 years) would be assessed as a medium- or high-risk individual. A 56-year-old with 15 pack-years, high BMI (e.g. 33), high smoke intensity (e.g. 40

Table 4
Performance of the HUNT Lung Cancer Model versus NLST criteria for lung cancer (LC) diagnosis within 6 years in the validation cohort (CONOR) of ever-smokers with complete data using as threshold the 16% quantile of risk of events in HUNT corresponding to a LC risk at least 1·75% in ~16 years or 0·64% in 6 years or ~15 points in the nomogram. Of the 45,117 ever-smokers, 1986 were picked by the NLST criteria. Sensitivity, specificity, PPV and NPV are calculated based on including all participants, 10,000, selected by the HUNT Lung Cancer Model.

	Participants with LC (N)	Participants without LC (N)	Participants total (N)	Predictive value
HUNT Lung Cancer Model criteria[a]	270	45 117	45 387	
Criteria positive	221 TP (2·21%)	9 779 FP (97·79%)	10 000	PPV 2·21%
Criteria negative	49 FN (0·14%)	35 338 TN (99·86%)	35 387	NPV 99·86%
Sensitivity	81·85%			
Specificity		78·31%		
NLST criteria				
Criteria positive	66 TP (0·66%)	9 934 FP (99·44%)	10 000[a]	PPV 0·66%
Criteria negative	204 FN (0·57%)	35 183 TN (99·43)	35 387	NPV 99·43%
Sensitivity	24·44%			
Specificity		77·98%		

FN = false negative; FP = false positive; NPV = negative predictive value; PPV = positive predictive value; TN = true negative; TP = true positive.
[a] Total criteria positive selected by the HUNT Lung Cancer Model includes the 1986 picked by the NLST.

cigarettes per day), no periodical or daily cough, and no indoor smoke exposure would be assigned a lower than cut-off risk and would not be eligible for screening (risk score 12.5, LC risk = 1·35% at 16 years and 0.61% at 6 years, Fig. 2a, b).

4. Discussion

Accurate risk prediction modelling is key for selecting individuals for LC screening by CT. Here we propose the HUNT Lung Cancer Model, a simple yet highly predictive, externally validated model employing seven clinical factors. Five of the seven were established previously (age, pack-years, smoking intensity, BMI, and quit time) (Ten Haaf et al., 2017), while two are novel (hours of indoor smoke exposure and daily cough during periods of the year). The model is applicable to ever-smokers of all ages.

The predictive performance of the HUNT Lung Cancer Model is relatively high compared to other externally validated models in the literature and restricted to the task of predicting the 6-year LC incidence in the validation cohort, the sensitivity and specificity was 81·9% and 78·3% respectively, and the AUC was 0·87 [0·85–0·89] (Tables 3 and 4). A direct comparison of performance is however only possible for the NLST criteria, since each model was derived from a different population. In a recent paper by Ten Haaf et al. (2017) reviewing predictive models, the models including PLCOm2012, Bach, and the Two-Stage Clonal Expansion model predicted 6-year LC incidence in the PLCO chest radiography arm with an AUC estimated between 0·77 and 0·8. In several recent validations studies of the PLCOm2012 model in various populations of ever-smokers, there were also good AUC values (AUC = 0·80–0·81) (Ten Haaf et al., 2017; Li et al., 2015; Weber et al., 2017) and the EPIC study validation exhibited an estimated AUC of 0·843; however, participants in all these studies had an almost double or triple median number of pack-years (Table 3) (Hoggart et al., 2012). While different prediction models are not directly comparable, due to various populations and study designs, the HUNT Lung Cancer Model had high performance in terms of both AUC and C-index, in addition to having the longest follow-up time and largest age-span of all studies to date.

We adopted a data-driven approach for performing the analysis and selecting the variables in the model. Methodologically, the analysis included a novel pipeline consisting of non-linear transformation of the continuous variables, testing for inclusion of second-order interaction terms, using multiple imputation to address the uncertainty from missing values, feature selection for determining the subset of important factors, estimating coefficients according to Rubin's rules, and bootstrapping for internal validation. The importance of non-linear transformation was seen in the analysis of LC risk in the HUNT population (Goodness-of-fit chi-squared statistic Supplementary Fig. S1 and Supplementary Table S2). This non-linear effect is also corroborated in the literature (Meiners et al., 2015). In contrast, the competing models, including the NLST criteria, often focus on a single age group (Table 2) and omit younger and older people. Importantly, among the LC cases in CONOR, 21.35% were younger than 55 and 18.41% were above 74 at baseline, indicating that the NLST age cut-offs by itself excludes almost 40% of LC in our cohort.

High smoking burden is a known strong risk factor for and was present in this study as well. However, to our knowledge, no prior model has been fit or tested in cohorts of light smokers of all ages (Table 2). In the HUNT2 and CONOR populations, median pack-years were 10·3 and 11·5, respectively, while 64% and 61% of ever-smokers who developed LC had smoked <30 pack-years (Figs. 3a and 4a).

Table 5
Accuracy of NLST versus the HUNT Lung Cancer Model for lung cancer (LC) diagnosis within 6 years, using the same number of screenings as NLST in the CONOR ever-smokers with complete data. As compared with NLST criteria, our model's criteria identified 103 vs 69 out of 270 cases showing an improved sensitivity (38.14% vs 25.6%, P = 0.0216) and positive predictive value (4.95% vs 3.3%, P < 0.000001), with the same specificity (95.61% vs 95.5%, P = 0.7321) and similar negative predictive value (99.6% vs. 99.5%, P = 0.95374).

Criteria[a]	Participants with LC (N)	Participants without LC (N)	Participants total (N)	Predictive value
NLST criteria	270	45,117	45,387	
Criteria positive	69 TP (3·3%)	2012 FP (96·7%)	2081	PPV 3·3%
Criteria negative	201 FN (0·5%)	43,105 TN (99·5%)	43,306	NPV 99·5%
Sensitivity	25·6%			
Specificity		95·5%		
HUNT Lung Cancer Model criteria				
Criteria positive	103 TP (4·95%)	1978 FP (95·05%)	2081	PPV 4·95%
Criteria negative	167 FN (0·4%)	43,139 TN (99·6%)	43,306	NPV 99·6%
Sensitivity	38·14%			
Specificity		95·61%		

FN = false negative; FP = false positive; NPV = negative predictive value; PPV = positive predictive value; TN = true negative; TP = true positive.
[a] NLST criteria for study entry included a history of cigarette smoking of at least 30 pack-years, age between 55 and 74 years and, for former smokers, cessation within the previous 15 years.

In line with others, we found that smoking the same total number of cigarettes over a long versus a short time span increases the risk of LC and is a significant risk predictor in the model (Table 2, Fig. 2a) (Vlaanderen et al., 2014). Interestingly, in a meta-analysis of 15 studies, this correlation was at best uncertain below 20 pack-years, however, we found that this effect also persists in a light smoker population (Vlaanderen et al., 2014).

The proposed model also found that smoking cessation is an important factor in which the risk decreases according to the logarithm of cessation time, in line with others (Fig. 2a) (Vlaanderen et al., 2014). In HUNT2, 29% of those developing LC were former smokers and 27% had quit <30 years previously (Fig. 3b). Imposing an empirical cutoff on the quit time as in the NLST study is probably not the best strategy.

Meta-analysis has confirmed the inverse proportional role of BMI in LC risk, but the biological basis for the protective effect of high BMI is not well understood (Duan et al., 2015). Low BMI was also found to be a negative predictor in the HUNT Lung Cancer Model. One explaining hypothesis could be that the biological factors that produce weight loss in smokers are a proxy for genetic susceptibility to LC.

Our analysis revealed two independent predictors that have not been used in previous models: daily cough and hours of exposure to smoke. Ever-smokers answering "yes" on the question "Do you cough daily during periods of the year?" had a statistically significant higher risk (Table 2). Daily cough is the only symptom of the ten included that was selected by the analysis (Table 1). Chronic or periodic cough is common and may be elicited by many non-cancer factors, including exposure to cigarette smoke and environmental pollution and a is key symptom in diseases such as chronic obstructive pulmonary disease (COPD), asthma, eosinophilic bronchitis, rhinosinusitis, pulmonary fibrosis bronchiectasis and even gastro-oesophageal reflux disease (Chung and Pavord, 2008; Kessler et al., 2011). In CONOR, 18·3% of those who did not develop LC had this symptom versus 34·1% of those who developed LC ($p < 10\text{-e4}$, Table S10), indicating how common this symptom is in the population. In current smokers it has been noted as a predictor of LC (Islam and Schottenfeld, 1994). Daily cough in ever-smokers could indicate early damage or frailty of the airways or be a symptom of an early cancer and clearly warrants further study.

Answering a high number of hours to the question "How long are you usually in a smoky room each day (hours)?" was a statistically significant independent risk predictor in ever-smokers, indicating that ever-smokers may also suffer damage from second-hand smoking of their own or other smokers´ and that the total smoke exposure matters (Table 2). Even in countries where indoor smoking has been banned, a clinician should therefore remember to ask this question when assessing LC risk for a current or a former smoker.

In line with several other studies we were unable to find significant clinical predictors for LC risk in never-smokers (Weber et al., 2017; Hoggart et al., 2012). Perhaps supplementing these studies with molecular markers could lead to identifying predictors for this subpopulation.

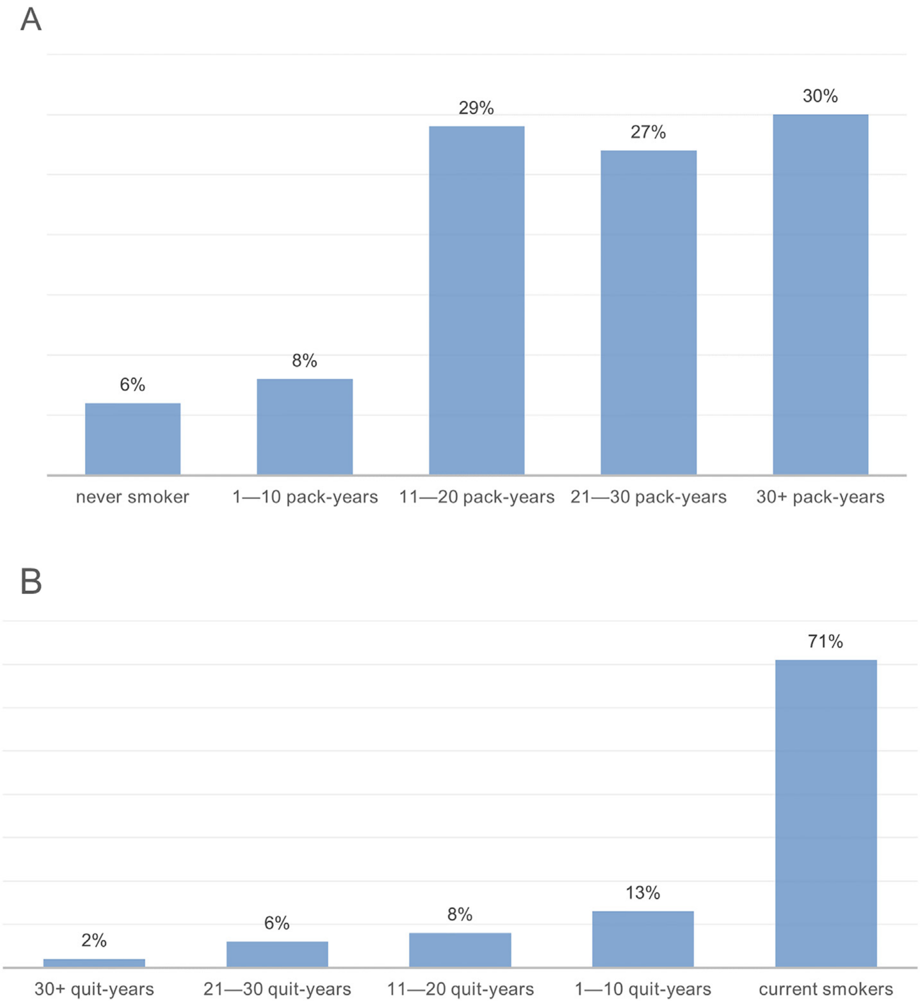

Fig. 3. Smoking status among lung cancer cases in HUNT2. A. Pack-years distribution at enrolment (not at diagnosis). Of importance, the majority were current smokers, and the 10, 20, and 30 pack-years groups all had a similar size, with 70% of those who developed cancer having smoked <30 pack-years at baseline. B. Distribution of current and former smokers at enrolment; 27% of those who developed lung cancer had a smoking quit time of <30 years.

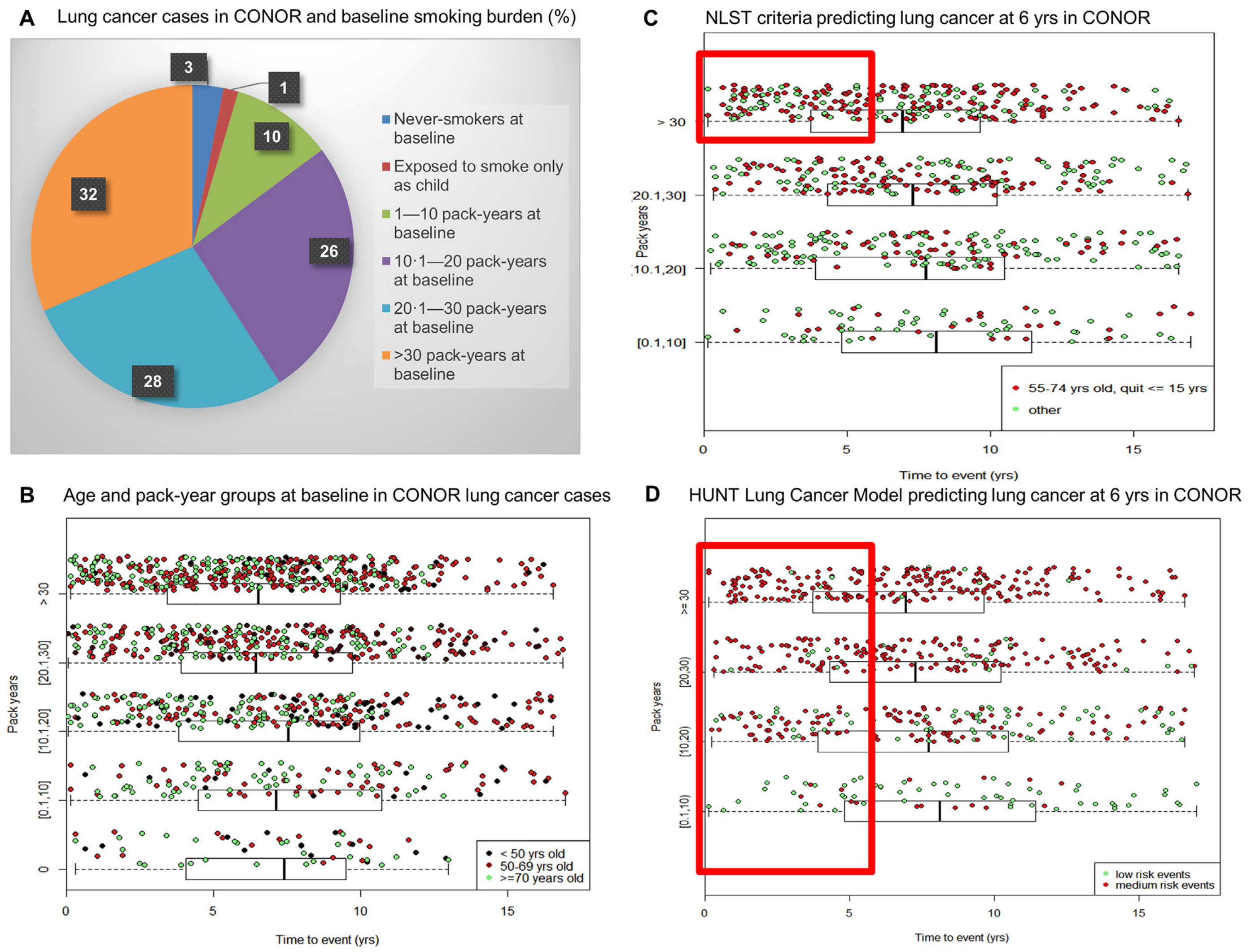

Fig. 4. Characteristics of the lung cancer events in the validation population (CONOR) with 0–20 years of follow-up. A. Overall distribution of smoke exposure in lung cancer cases in CONOR ($n = 709$). B. Lung cancer event appearance in ever-smokers ($n = 675$) in the CONOR population after baseline (x-axis) according to age groups (colour code) and pack-years (y-axis) showing that all age groups are represented in all pack-year groups. In the three vertical boxplot (B, C and D) median time to diagnosis is not significantly different between pack-year groups (not shown). C. Lung cancer event appearance in the CONOR population (all dots) after baseline (x-axis) and pack-years (y-axis) according to NLST criteria (red dots; >30 pack-years, 55–74 years of age and <15 years quit time). Within 6 years, less than one third of the total cases would be included in the NLST screening (cases in red within the red quadrant). D. HUNT Lung Cancer Model applied to the CONOR population after baseline registration. Lung cancer event appearance in the CONOR population after baseline (x-axis) and pack-years (y-axis). Red dots are lung cancer cases predicted using the model according to medium- plus high-risk groups in HUNT (Fig. 2b) corresponding to the 16% quantile of the risk of events in HUNT, equalled to a risk of 1·75% for developing LC within 16 years (>15 points in the nomogram). Based on this threshold, 221/270 LC events within 6 years (red dots within the red square) and 527/675 events within ~20 years were correctly predicted. More specifically, using this threshold, one would need to examine 9998 out of 45,387 (22%) ever-smokers to identify 82% of future events in a 6-years period or 78% of future events in a 20-year period (median 11·6 years).

Sex was not selected by the backwards feature selection process, but its inclusion was enforced in the model as a commonly used factor in risk models for cancer (Table 2, Supplementary Appendix). Nevertheless, its coefficient as well as its interaction with smoking variables (Supplementary Tables S7–S8) was not found to be significant, in accordance with an accumulated body of research showing that women exposed to first- or second-hand cigarette smoke have the same risk as men for developing LC (De Matteis et al., 2013).

A proxy for genetic susceptibility is included among the 36 variables in the form of the question, “Do you have or have you ever had cancer?” This variable was significant in the univariate but not the multivariate analysis (Supplementary Table S1a and b). Family history of LC is a known risk factor and a variable included in some prediction models (Weber et al., 2017; Marcus et al., 2015). Family history of LC or cancer in general, is a variable that is often hard to accurately obtain as some people may not know details of their family history for various reasons. A model where all the variables can be obtained reliably and consistently therefore has a clinical advantage.

Educational status is a known predictor for LC and may also be a proxy indicator of social and economic status, including smoking behaviour. Norway has had a public educational system and an egalitarian social democratic system since the 1950s that may explain why this variable was not significant in this study. In addition, smoking behaviour is included in the model, which may explain the absence of the effect of the educational status when controlling for all other factors.

Regarding transferability to other populations, we do not know if people from different social or ethnic background would need a different model. In our validation population there were 1.4% non-Europeans, there were populations from the capital and big cities as well as a large population from rural areas, there were people born before World War II and up to 1976 (Naess et al., 2008). In many ways the population is quite heterogenous, which is a strength of the model.

Some limitations of the study are the lack of some known LC risk predictors, including history of COPD, occupational exposure to asbestos or radon, and heredity. Regarding COPD, among the 36 chosen variables, several proxies indicate a chronic lung condition, such as self-perception of health, ever-asthma, ever-asthma medication, wheezing last year, cough daily, cough with phlegm, dyspnoea last year, and medication daily last year. These factors were significant only in the univariate analysis, except daily cough, which was selected in the model (Table 1, Supplementary Table S1b).

Radon exposure, pollution, and other environmental factors are important, but the low LC rate of <6% never-smokers indicates a minor contribution of non–smoke-related factors. Regarding asbestos exposure, we do not know whether our model has the same accuracy in predicting LC in heavily asbestos-exposed populations because asbestos inhalation highly potentiates the risk for LC, especially in combination with cigarette smoking. Because of abolition of asbestos in most industrialized countries in the 1980s, individuals heavily exposed to asbestos are a minority, but this issue persists in some countries where asbestos is widely used (Roe and Stella, 2015). In other areas of the world with high environmental or occupational carcinogen exposure, these factors would need to be addressed in future risk models.

This study, based on data from large, prospective, population-based studies in Norway with a long follow-up time, identified a new, highly predictive risk model, the HUNT Lung Cancer Model. This is the first risk prediction model developed and tested in a light-smoker population of all ages. Our research proves the model's effectiveness in selecting high-risk individuals for LC screening. A nomogram and an online calculator facilitate the calculation of 5-, 10-, 15-year as well as the 6- and 16-year risk of LC diagnosis respectively. The innovation, compared with previous models, is the establishment of seven significant clinical predictors by feature selection, including newly identified predictors, such as daily cough and hours exposed to smoke. Furthermore, the model's simplicity and its applicability to all ages and even light smokers represents an important improvement to the NLST criteria, for effective CT screening.

Declaration of Interests

The authors have no conflicts of interest.

Funding

The Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (46056639), and Bonnie J. Addario Lung Cancer Foundation and International Association for the Study of Lung Cancer (IASLC) Joint Award 2016.

Author Contributions

Conception and design: Oluf Dimitri Røe.

Collection and assembly of data: Kristian Hveem, Arnulf Langhammer.

Data analysis and interpretation: Maria Markaki, Vincenzo Lagani, Ioannis Tsamardinos, Oluf Dimitri Røe.

Manuscript writing: Maria Markaki, Oluf Dimitri Røe, Vincenzo Lagani, Ioannis Tsamardinos.

Final approval of manuscript: All authors.

Acknowledgements

We want to thank all the participants in the HUNT and CONOR population studies, the HUNT research staff, Michael Tsagris, PhD, for useful discussions, Iordanis Xanthopoulos, MSc, for developing the online calculator, Alex Andreou and Theo Theoharis for language revision.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2018.03.027.

References

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., et al., 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 351, h5527.

Chung, K.F., Pavord, I.D., 2008. Prevalence, pathogenesis, and causes of chronic cough. Lancet 371 (9621), 1364–1374.

De Matteis, S., Consonni, D., Pesatori, A.C., et al., 2013. Are women who smoke at higher risk for lung cancer than men who smoke? Am. J. Epidemiol. 177 (7), 601–612.

Duan, P., Hu, C., Quan, C., et al., 2015. Body mass index and risk of lung cancer: systematic review and dose-response meta-analysis. Sci. Rep. 5, 16938.

Harrell, F.E., 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York.

Heymans, M.W., van Buuren, S., Knol, D.L., van Mechelen, W., de Vet, H.C., 2007. Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med. Res. Methodol. 7, 33.

Hoggart, C., Brennan, P., Tjonneland, A., et al., 2012. A risk model for lung cancer incidence. Cancer Prev. Res. (Phila.) 5 (6), 834–846.

HUNT Variables, 2018. https://www.ntnu.no/hunt/variabler.

Islam, S.S., Schottenfeld, D., 1994. Declining FEV1 and chronic productive cough in cigarette smokers: a 25-year prospective study of lung cancer incidence in Tecumseh, Michigan. Cancer Epidemiol. Biomarkers Prev. 3 (4), 289–298.

Katki, H.A., Kovalchik, S.A., Berg, C.D., Cheung, L.C., Chaturvedi, A.K., 2016. Development and validation of risk models to select ever-smokers for CT lung cancer screening. JAMA 315 (21), 2300–2311.

Kessler, R., Partridge, M.R., Miravitlles, M., et al., 2011. Symptom variability in patients with severe COPD: a pan-European cross-sectional study. Eur. Respir. J. 37 (2), 264–272.

Krokstad, S., Langhammer, A., Hveem, K., et al., 2013. Cohort profile: the HUNT study, Norway. Int. J. Epidemiol. 42 (4), 968–977.

Li, K., Husing, A., Sookthai, D., et al., 2015. Selecting high-risk individuals for lung cancer screening: a prospective evaluation of existing risk models and eligibility criteria in the German EPIC cohort. Cancer Prev. Res. (Phila.) 8 (9), 777–785.

Marcus, M.W., Chen, Y., Raji, O.Y., Duffy, S.W., Field, J.K., 2015. LLPi: liverpool lung project risk prediction model for lung cancer incidence. Cancer Prev. Res. (Phila.) 8 (6), 570–575.

Marshall, A., Altman, D.G., Holder, R.L., Royston, P., 2009. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med. Res. Methodol. 9, 57.

Meiners, S., Eickelberg, O., Konigshoff, M., 2015. Hallmarks of the ageing lung. Eur. Respir. J. 45 (3), 807–827.

Moons, K.G., Altman, D.G., Reitsma, J.B., Collins, G.S., 2015. Transparent reporting of a multivariate prediction model for individual prognosis or development I. New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the TRIPOD statement. Adv. Anat. Pathol. 22 (5), 303–305.

Muller, D.C., Johansson, M., Brennan, P., 2017. Lung cancer risk prediction model incorporating lung function: development and validation in the UK Biobank Prospective Cohort Study. J. Clin. Oncol. 35 (8), 861–869.

Naess, O., Sogaard, A.J., Arnesen, E., et al., 2008. Cohort profile: cohort of Norway (CONOR). Int. J. Epidemiol. 37 (3), 481–485.

National Lung Screening Trial Research Team, Aberle, D.R., Adams, A.M., et al., 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. 365 (5), 395–409.

Oudkerk, M., Devaraj, A., Vliegenthart, R., et al., 2017. European position statement on lung cancer screening. Lancet Oncol. 18 (12), e754–66.

Patz Jr., E.F., Pinsky, P., Gatsonis, C., et al., 2014. Overdiagnosis in low-dose computed tomography screening for lung cancer. JAMA Intern. Med. 174 (2), 269–274.

Pepe, M.S., Feng, Z., Huang, Y., et al., 2008. Integrating the predictiveness of a marker with its performance as a classifier. Am. J. Epidemiol. 167 (3), 362–368.

Pinsky, P.F., Berg, C.D., 2012. 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. 19 (3), 154–156.

Raji, O.Y., Duffy, S.W., Agbaje OF, et al., 2012. Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. Ann. Intern. Med. 157 (4), 242–250.

Rampinelli, C., De Marco, P., Origgi, D., et al., 2017. Exposure to low dose computed tomography for lung cancer screening and risk of cancer: secondary analysis of trial data and risk-benefit analysis. BMJ 356, j347.

Roe, O.D., Stella, G.M., 2015. Malignant pleural mesothelioma: history, controversy and future of a manmade epidemic. Eur. Respir. Rev. 24 (135), 115–131.

Royston, P., Altman, D.G., 2013. External validation of a Cox prognostic model: principles and methods. BMC Med. Res. Methodol. 13, 33.

Sogaard, A.J., Selmer, R., Bjertness, E., Thelle, D., 2004. The Oslo Health Study: the impact of self-selection in a large, population-based survey. Int. J. Equity Health 3 (1), 3.

Tammemagi, M.C., Katki, H.A., Hocking, W.G., et al., 2013. Selection criteria for lung-cancer screening. N. Engl. J. Med. 368 (8), 728–736.

Ten Haaf, K., Jeon, J., Tammemagi, M.C., et al., 2017. Risk prediction models for selection of lung cancer screening candidates: a retrospective validation study. PLoS Med. 14 (4), e1002277.

Torre, L.A., Siegel, R.L., Ward, E.M., Jemal, A., 2016. Global cancer incidence and mortality rates and trends—an update. Cancer Epidemiol. Biomark. Prev. 25 (1), 16–27.

van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R. J. Stat. Softw. 45 (3), 1–67.

Vlaanderen, J., Portengen, L., Schuz, J., et al., 2014. Effect modification of the association of cumulative exposure and cancer risk by intensity of exposure and time since exposure cessation: a flexible method applied to cigarette smoking and lung cancer in the SYNERGY Study. Am. J. Epidemiol. 179 (3), 290–298.

Weber, M., Yap, S., Goldsbury, D., et al., 2017. Identifying high risk individuals for targeted lung cancer screening: independent validation of the PLCOm2012 risk prediction tool. Int. J. Cancer 141 (2), 242–253.

Wu, X., Wen, C.P., Ye, Y., et al., 2016. Personalized risk assessment in never, light, and heavy smokers in a prospective cohort in Taiwan. Sci. Rep. 6, 36482.