EBioMedicine xxx (2018) xxx–xxx

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 xxxx

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

Please cite this article as: Markaki, M., et al., A Validated Clinical Risk Prediction Model for Lung Cancer in Smokers of All Ages and Exposure Types: A HUNT Study, EBioMedicine (2018), https://doi.org/10.1016/j.ebiom.2018.03.027