Peer Review File

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

Reviewer A

This retrospective cohort study examined patients in Southern Denmark suspected of lung cancer (LC) between 2009 and 2018. The primary outcome was the diagnosis of LC, defined by the ICD-10 code C34. The study included 38,944 patients, with 29% LC and 71% non-LC. Data sources included LC diagnosis, comorbidity, prescription medication, smoking history, number of consultations, C-Reactive Protein (CRP) rapid tests, and routine blood sample analysis.

It found that 45% of the cohort were not registered with any comorbidity-codes, with conditions like diabetes, vascular disease, myocardial infarction, and metastatic solid tumors present in a higher proportion of lung cancer patients than in the non-LC cohort. The study also found that 27% of the total cohort were missing in the dataset, with all drugs prescribed to a higher proportion of lung cancer patients than non-LC patients.

LC patients were more likely to be smokers, have comorbidities associated with atherosclerosis and diabetes, and have a higher rate of prescribed prescription medicine. LC patients had significantly higher values of white blood cells, platelets, calcium, CRP, LDH, and alkaline phosphatase, and lower values of hemoglobin, albumin, ALAT, and creatinine.

This study aimed to describe the process of data collection, handling of missing data, and testing for isolated associations between various risk factors and lung cancer (LC).

I have the following comments:

1) The study shows very well how to deal with large databases. However, it is unclear to me why the authors did not consider the variables of the known scores for LC patients. These would be the Thoracoscore, Epithor, Eurloung 2, and the simplified Eurolung 2.

Reply 1: We agree with reviewer A that it would be ideal and relevant to include these scores to gain further insight on the lung cancer population. Unfortunately, they are not routinely collected or described in Denmark, and not reported in the Danish Lung Cancer Registry. Furthermore, data necessary for calculations retrospectively are missing. For instance, the Thorascore requires information on ASA classification, dyspnea score, and priority of surgery among others, of which none of these are routinely collected on Danish lung cancer patients. All scoring systems would be relevant if our aim was to predict an outcome within the lung cancer population (e.g. postoperative 30-day mortality). However, since our aim is to predict lung cancer from non-lung cancer patients, these score would not be included as variables, since they only apply to the lung cancer population.

2) Continuous data should be examined with the two tailed Student's t-test, categorical variables based on Fisher's exact test.

Reply 2: We appreciate the comment from reviewer A. We agree that we would use the student's t-test if data were normally distributed, but since none of the continuous variables followed a normal distribution the non-parametric Wilcoxon signed-rank test was applied. Since we are dealing with a large sample size, we used the chi-squared method for categorical variables. We agree that it would be relevant to use the Fisher's exact test if dealing with very small sample sizes (e.g. frequencies <5), but due to the large sample sizes the chi-squared method was the choice of use.

3) Afterwards significant variables should be included into a multivariate logistic regression analysis.

Reply 3: We agree with reviewer A, that this is the general approach when dealing with conventional statistics. The aim of this article was however to describe the different datasets and the groundwork to combine them into one dataset at the end, which would be use for prediction analyses. We are currently working on a prediction model that aims to predict lung cancer based on the combined dataset. We are using different machine learning methods, of which logistic regression is one among them. In prediction models we include all variables, no matter if they are significant or not from the inferential statistics, because small differences might matter in a non-linear model like many machine learning models. We appreciate the suggestion from reviewer A, but have chosen not to include a logistic regression model in this article. This is because we fear that it will confuse the overall message of the article, which is to describe the two cohorts, deal with large datasets with missing data and prepare them for prediction analyses.

4) The performance of the different scoring systems should be assessed by evaluation of calibration and discrimination. Calibration compares the observed mortality with that predicted by the model within severity strata. The most accepted method for measuring calibration is the goodnessof-fit statistic analysis, which uses a chi-square test (Hosmer—Lemeshow x2 statistics). Small x2 values and high corresponding p values indicate good calibration. Discrimination, or the ability of a scoring system to distinguish between survival and mortality, should be measured by the area under the receiver operating characteristic (ROC) curve. The ROC curve shows the relation between sensitivity and specificity.

Reply 4: We agree with reviewer A, but as mentioned in reply 1, we are not able to collect the variables necessary to calculate these scores.

Reviewer B

1) This is a retrospective review of more than 40,000 patients in the Southern Danish Registry who were initially coded as being evaluated for concerns of lung cancer, and later identified as having lung cancer or a non-lung cancer diagnosis. The authors evaluate the incidence and proportion of lung cancer and non-lung cancer diagnoses over time, and correlate comorbid conditions, prescription medications and laboratory findings with these diagnoses. Overall, this is highly descriptive with some minor novel findings, some of which have questionable clinical import (including much of the blood test levels as described). The most potentially interesting finding was an increase in the proportion of early stage lung cancers over time. This could be much more impactful if the authors were able to correlate increases in early-stage lung cancer diagnosis with reductions in time to evaluation and/or treatment in this cohort. Additional major and minor critiques are included below.

Major Critiques:

1) Figure 3: It is not clear what the large colored stars are supposed to denote. If statistical significance, please note this within the figure legend and provide more common use of *, ** or *** (depending on the level of statistical significance).

Reply 1: We appreciate the comment and have now modified the figure legends for figure 3, 5 and 6.
Changes in the text: Please see the modified legends (p.14, line 524-538).

2) The percentage of patients (lung cancer and non-lung cancer) identified as having comorbid conditions is quite low. This does not match existing published data. Additionally, the authors' own cohort identifies a higher percentage (25-30%) as receiving medications for COPD, while only 12-13% were identified as having this comorbidity. It seems this most likely represents incomplete data or inability to accurately extract this data from the database used. Can the authors please comment on this and consider whether this data, if it cannot be confirmed accurate, should be published.

Reply 2: We agree with reviewer B that 62% of the lung cancer cohort and 65% in the non-lung cancer cohort without a comorbid condition included in the CCI are large proportions. This is also mentioned in the discussion at p.10, line 354, and believed to be partly due to the Danish registration of ICD-codes, which are only registered at a hospital level. This explains why only 12-13% are registered with COPD in this study, while 25-30% receive medication for COPD. The 25-30% include patients followed with COPD in general practice, whereas the COPD ICD-10 codes include only patients followed with COPD at a hospital level. If it becomes possible to include data from general practice, we will gain insight to this gap.
Changes in the text: We have now emphasized this aspect the discussion (p. 11, line 421).

3) I'm not sure that any clinically relevant information can be gleaned by comparing lung cancer and

non-lung cancer consultations by the actual number of consultations (particularly as there are multiple non-significant visit numbers). The authors should consider evaluation of certain ranges of consultations, which seem much more appropriate than comparing, for instance, 17, 16, 15, 20 (etc.) consults for each separately.

Reply 3: We thank reviewer B for pointing this out, and have now modified into only two comparisons. One is the comparison of the proportion of patients with actual visits (>0) versus no visits in both groups. Another is between the intervals 1-4 visits compared to >4 visits. This cutoff was chosen based on a clinical approximation of 1-4 visits in 6 months being a relatively normal rate, and >4 visits a rather high visit rate on this cohort of patients.

Changes in the text: We applied these thresholds for the CRP rapid test results as well. We have modified the relevant text in the methods (p.6, line 181) and results (p.8, line 296).

4) While I appreciate the attempt to look at blood markers in these populations (lung cancer and non-lung cancer), many of the statistically significant data is clearly not clinically significant. For instance, INR is showed as having a p<0.001, but median vales (and ranges) are exactly the same for both groups. The same can be said for numerous others which have little or no difference, and certainly not what would be considered a clinically significant difference (pretty much all of them-possibly excluding neutrophils and leukocytes which showed a small difference). I am not sure the results are relevant to be presented or, if presented, should be supplemental and it should be clearly described that the majority of the differences would not be considered of any clinical significance.

Reply 4: We agree with reviewer B that most median values are close, and that differences are mostly minor within the reference interval, not necessarily considered to be clinically relevant. However, even small differences in parameters might lead to improved discrimination resulting in a better performance in a prediction model. While traditional statistical models aim to inferring causality or associations between variables, machine learning aims to make accurate predictions on large complex dataset with non-linear interactions. Therefore, we believe it is relevant to report these results, even though we agree that they may not be clinically notable. We will reassess this decision if the reviewers find it to be essential.

Changes in the text: We have modified the results (p. 9, line 312) and discussion (p.12, line 429).

We appreciate the comment on the significance of the INR results. The analysis is based on the Wilcoxon signed rank test, which is a sum test and not a median test. Therefore it is possible to obtain a significant results based on the same median (as for INR). We do however understand if it can create confusion, and will be open to consider stating an explanation e.g. as a footnote to the table 1 if relevant.

5) Of high interest is the finding that the proportion of lung cancers diagnosed at an early stage increased over the studied period of time. The authors surmise that this may be due to reduction in delay of diagnosis through implementation of the Patient Pathways (2007). It would be highly impactful if the authors were able to show a corresponding decrease in time from consultation to diagnosis in the portion of the cohort with lung cancer over this period of time (which, given their other data measures, would likely be measurable).

Reply 5: We thank reviewer B for this comments and agree that the increase in proportion of early stage lung cancers over time is interesting, and would be relevant to link to time to from examination to diagnosis. However, the Danish Lung Cancer Registry has decided that the date of initiation in the LC fast-track clinics also marks the diagnosis date. Consequently, it is not possible to investigate the actual number of days between initiation of the LC fast-track examinations and the date of final diagnosis. One rationale for this decision is that we anticipate minimal variations among lung cancer patients within this time interval, given that fast-track lung cancer clinics have mandated a maximum of 30 days between the initial examination and the beginning of treatment since 2007.

A more evident explanation for the increased number of low-stage patients throughout the study period is that more patients are being referred to lung cancer fast-track clinics, leading to an increase in CT scans. Additionally, in 2007, there was an expansion allowing general practitioners to directly refer patients for CT scans in cases of vague symptoms suggestive of cancer. The increased frequency of CT scans has been shown to result in the detection of more low-stage lung cancer cases.

Changes in the text: We have elaborated on this explanation in the discussion (p.11, line 393).

Minor:

6) Figure 1: graphic representation of data should probably have 7 "green people" to represent non-lung cancer cases and 3 "red people" to represent lung cancer cases, as these are more closely aligned with the data presented by the authors

Reply 6: We appreciate this suggestion and have now modified Figure 1 according to the suggested changes.

7) p.6: please define "esophagus-ventricle cancers"

Reply 7: Esophagus cancer was defined by the two ICD-10 code C15* Malignant neoplasm of esophagus, and ventricle cancer by C16* Malignant neoplasm of stomach. We have now modified this to "esophagus-stomach cancers" and attached a table in the supplementary file (supplementary file 2), listing cancer locations and corresponding ICD-10 codes included in "other malignancies".

Changes in the text: We changed the text (p.7, line 253) and Supplementary file 2.

8) Numerous minor grammatical errors in English language throughout – these should be corrected through use of a grammatical correction tool (such as Microsoft word) and/or through proof reading by someone with extensive knowledge of English grammatical structure- Specific example on p.7: "Exploration of missing data revealed that 10% of the total cohort did not have any registration of neither consultation nor CRP rapid test within the 6-month interval before the index date." – This should read "Exploration of missing data revealed that 10% of the total cohort did not have any registration of either consultation of CRP rapid test within the 6-month interval before the index date." P.7: "lymfocyctes" should be "lymphocytes", "albumin" should be "albumin"

Reply 8: We thank reviewer B for noting these grammatical errors and have now corrected them.

Changes in the text: We changed the text (p.8, line 292) (p.9, line 315).

9) Figures 4, 5 and 6A: Please correct the multiple misspelled diseases (lymphoma, prostate, esophagus, head and neck…" and again, consider definition of "esophagus-ventricle"

Reply 9: We once again thank the reviewer for noting these errors and have now corrected them.

Changes in the text: We have modified the text in Figure 4, 5 and 6.

10) Discussion: I would argue that the following statement is not accurate: "Despite the similarities, we were able to distinguish LC patients based on factors such as age, smoking status, comorbidities and laboratory results." As this study shows considerable overlap, and was not designed to create clinical/biomarkers for differentiation, LC patients were not able to be distinguished based on these characteristics. However, one may say that there were statistically-significant differences within these parameters between the two groups.

Reply 10: We appreciate the feedback on the conclusion provided by reviewer C and have omitted the mentioned sentence, as we recognize that it presents an overly clear-cut interpretation of the results. Considering the article's objective, we still deem it relevant to describe clinical/biomarkers for differentiation. We agree that, based on the conventional statistical tests presented in this manuscript, we cannot definitively distinguish between LC and non-LC patients, and acknowledge that even minor differences in results can attain statistical significance in large cohorts without necessarily carrying clinical significance. Nevertheless, we find these subtle distinctions relevant, as they align with existing literature. Furthermore, when integrated into a high-dimensional model capable of handling complex distributions, these nuances may contribute to a predictive model.

Changes in the text: We have changed the conclusion and abstract with this feedback in mind (p.2, line 44, p.13, line 492).

Reviewer C

In this manuscript, Henriksen et al. report on various risk factors associated with lung cancer in a large Danish chort (~40,000 patients). I applaud the authors for their efforts in pursuing this venue. I have the below comments:

Major:

1) Main critique is how this work changes our everyday practice. As expected from such a large dataset,

there are various findings with significant p-values. However, the actual difference is rather subtle, and I am not sure how the findings will lead to any change. As the authors state in the introduction, what we lack is a refined screening criteria. Based on the findings, what is the authors' main conclusion? I wonder if a more rigorous statistical method would allow the authors to parse out which factors are truly significant and hence can be potentially utilized as a screening criteria.

Reply 1: We thank reviewer C for this relevant comment. As mentioned earlier, we acknowledge that even minor differences in results can attain statistical significance in large cohorts without necessarily carrying clinical significance. To account for multiple testing, we have changed our level of significance to p<0.01 (compared to 0.05). This changes the number of significant attributes, and we have corrected these changes throughout the manuscript.

Changes in the text: Please see the threshold (p. 6, line 220)

Since the last submission of the manuscript, we have finished the annotation of symptoms, familial predisposition data and relevant exposures on 5,587 patients. These patients were part of the 9,940 patients with complete results and adds another layer to the level of details. We have compared the variables for LC and non-LC patients, as well as compared the proportion of lung cancer patients with low (I-II) and high (III-IV) stage disease for each specific attribute. **Changes in the text:** This has resulted in a new figure (Figure 8) and additional paragraphs in the methods (p.6, line 198), results (p.9, line 319) and discussion section (p.12, line 446).

2) As somewhat expected in a large database study, there are many missing data (45% missing comorbidity information, for example). While the authors explain the "reduced cohort" with the complete data, it is unclear if data analysis on this cohort was performed.

Reply 2: We appreciate the comment from reviewer C regarding sub analyses on the reduced cohort with complete information. As mentioned in reply 1, we have now added data regarding symptoms, familial predispositions and relevant exposures. Consequently, the cohort with complete data is reduced to 5,587 individuals. We have performed sub analysis on this cohort, to investigate if the trends found in the initial datasets were persistent when reducing to this minor cohort.

Changes in the text: We have included this sub analysis supplementary file 4, and changed the results (p.9, line 334).

3) Lack of smoking history is a major criticism in my mind as this is a screening criteria in the Unites States.

Reply 3: We thank reviewer C for pointing out this important issue and agree that lack of smoking history on non-lung cancer patients is a major challenge in general.

Since this is a retrospective study, it was not feasible to obtain smoking data based on prospective questionnaires. In the NLST-trial in the United States, patients were enrolled based on mail, community outreach and mass media. In Denmark, we have only a few population surveys and none included substantial information on the population included in this study (patients from the Region of Southern Denmark). Therefore, we manually annotated smoking status from free text from the electronic health records. Sixty percentage of the cohort had usable information in the subfield relating to smoking status. We are currently working on a larger project that deals with obtaining risk factors from the general population in the Region of Southern Denmark by questionnaires (smoking, alcohol, exercise etc.). We agree that it would be relevant to incorporate such detailed information on risk factors such as smoking in the future, and have added a paragraph mentioning this aspect in the discussion.

Changes in the text: Please see the added paragraph (p.13, line 474).

4) For the general audience, it would be nice to the current lung cancer screening criteria in Denmark. Does one exist?

Reply 4: The current status in Denmark is that it has been decided on a political level to initiate a screening pilot study. The structure and magnitude of this pilot is still unknown but should be clarified within the next year.

Changes in the text: This is now mentioned this in the introduction (p.3, line 80).

5) I would potentially be interested in a separate sub-group analysis if the numbers allow. One is an analysis excluding the patients without any symptoms. This would be the true "screening" population. Another is an analysis focusing just on the never smokers as this group represents a population who would never be screened and hence in need of novel risk factors.

Reply 5: We thank reviewer C for this interesting suggestion. As mentioned in reply 1, we have now added information on symptoms etc. and are therefore able to investigate this matter further. The cohort with complete information from all datasets (including symptoms etc.) consists of 5,587 patients. In order to investigate this "true screening population" we have excluded patients with symptoms leading to referral for the lung cancer fast-track clinics (hemoptysis, pneumonia, cough, dyspnea, fever, weight loss, fatigue). Furthermore, we have excluded non-smokers, as well as individuals outside the age interval 50-80 years old. This resulted in 589 individuals that would actually be eligible for screening (10.5% of the 5,587 patients). 238 (40.4%) of these had LC and 351 (59.6%) did not.

If focusing on the subgroup of never smokers without symptoms leading to referral for the LC fast-track clinics, regardless of the age interval, this results in only 243 individuals (4.4% of the 5,587 patients). 30 (12.4%) of these had LC and 213 (87.7%) did not.

Since both of these two groups are considerable small, we have not continued with sub-group analyses on these sub cohorts as we did for the reduced cohort with complete results.

While we acknowledge the interest in considering both the true screening and never-screened populations, our study focuses on a specific cohort—patients referred for examination on suspicion of lung cancer. This group, mostly characterized by symptoms at referral, represents a high-risk population and differs from the general practice's true screening population. Despite this distinction, the insights provided in this article concerning data sources, data handling, and descriptive statistics hold value for future research conducted on populations with lower risk. We are currently in the process of testing a lung cancer prediction model using retrospective data from this high-risk cohort. We aim to evaluate its performance in subsequent assessments within populations characterized by lower risks, including patients under COPD monitoring at a hospital level or within general practice.

Minor:

6) This is just my personal interest, but neutrophil-to-lymphocyte (NLR) has been investigated heavily in other cancers. Since the authors have the data already, this may be interesting to look into.

Reply 6: We thank reviewer C for this interesting suggestion. The NLR has now been added to Table 1, and shows a NLR of 3.4 for the LC group compared to 2.6 for the non-lung cancer group ($p<0.001$). This pathological value for the LC group is interesting and have now been added to the results and discussion.

Changes in the text: Please see the addition to the results (p.9, line 316) and discussion (p.12, line 440).

7) I believe lines 310-311is a typo? (one group should be non-LC)

Reply 7: We appreciate the comment and have now corrected the mistake.

Changes in the text: Please see the corrected text (p.8, line 296).