Peer Review File

Article information: https://dx.doi.org/10.21037/tlcr-24-241

Reviewer A

This manuscript presents results from a population-based cohort study of individuals diagnosed with lung cancer in the UK, using a primary-care database. Although this is an important topic, I have concerns that dampen my certainty in this work. Major concerns are outlined below, followed by minor comments.

Major concerns:
Comment 1:
1) Additional clarity is needed from a methodological standpoint regarding why the study data were limited to the years 2000 to 2021, rather than an earlier timepoint? Based on the introduction, as well as my background knowledge on lung cancer screening, I am unsure why this timepoint was selected as the starting point of the cohort study.

Reply 1: Thank you for your comment. The study start was set as the year 2000 as this is when there is reliable data for both databases.

Comment 2:
2) Related to the study timepoints, an explanation is needed for why the data were only collected through 2021 and 2019 (instead of 2022, if the new screening program began in 2023)? Further, an explanation for why the data from CPRD GOLD were collected through December 2021 while the data from Aurum were collected through December 2019.

Reply 2: Thank you for your comment. We studied until 2019 in CPRD Aurum as there were problems with data quality of this database post 2019 that have only much later been resolved. Therefore, when the study was performed a study period of 2000-2019 was set. Although the study period for CPRD GOLD was 2000-2021 we wanted to still replicate our analyses in CPRD Aurum to show the robustness of our results. When the study was performed, we only had approval and access to data up to the end of 2021, so we could only perform the analysis up to this date for CPRD GOLD.

Changes in the text: None.

Comment 3:
3) Throughout the introduction and discussion, lung cancer risk factors such as smoking, environmental exposures (e.g., radon), and occupational chemicals are highlighted. However, it appears as though the relative contribution of these factors cannot be accounted for within the data. I understand that it may be difficult to gather additional and/or complementary data to help gain a better understanding of the contribution of these factors to the lung cancer diagnoses,

but I encourage the authors to explore options and include further detail if possible. The inclusion of smoking pack-years and years-quit would be especially helpful. Co-use of other smoking/tobacco products would also be helpful, if available.

Reply 3: We thank you for your comment. We agree with the reviewer further details on smoking would be useful. However, smoking is a difficult demographic to extract from primary care data due to its high level of missingness and differences in how it is recorded. Furthermore, CPRD specifically warn "*some health information such as smoking status, BMI, or ethnicity data may only be recorded when this is relevant to the patient's health condition, potentially creating bias in the patterns of completeness...*" (source: https://www.cprd.com/introduction-cprdhttps://www.cprd.com/introduction-cprd). Based on this, we have not included further details on smoking in this manuscript. However, we have added this valid point to the limitations of the manuscript. It is worth mentioning that our study does not aim to study the association between smoking and lung cancer, which is well proven and out of scope for this work.

Changes in the text: Page 17, Line 433-435 "Although not possible in CPRD due to missingness which could create biases in results, future research using alternative databases could be used to quantify the relative contribution of smoking in 'pack-years', alongside alternative discussed risk factors."

Comment 4:
4) In the abstract, introduction, and discussion, the authors state that the study is important because of a new screening program that was introduced in 2023. Although it would be helpful to establish a baseline number of people who meet the criteria (ages 55 to 74 with a documented smoking history) who have been screened with LDCT, this does not appear to be assessed in the current study.

Reply 4: Thank you for your comment – we agree this is an excellent proposed addition to the text. However, it must be noted that LDCT is a new screening programme that is still being rolled out, so, as described below, we refer to the predicted figures issued by the UK Government.
Changes in the text: Page 4 Line 99-101. "When the rollout is complete, it is estimated 325,000 people will become newly eligible for a scan every year"

Comment 5:
5) Similarly, I do not see where the study results report on lung cancer diagnoses and outcomes for patients who received LDCT screening based on a risk assessment versus diagnostic testing, incidental findings from another test, and other paths that may result in a lung cancer diagnosis. I highly encourage the authors to address both the screening rate and diagnoses resulting from screening (and believe the data should be available).

Reply 5: Thank you for your suggestion. LDCT screening has only been implemented since 2023 whereas our study end date was December 2021 Therefore, we cannot

differentiate the study outcomes for people who received LDCT screening versus diagnostics/incidental findings, as we do not have a population of these patients within the study period. However, we have added this to the discussion for future work as the reviewer makes a very valid suggestion.

Changes in the text:
Page 16, Line 397-399:
The introduction of CT screening may increase the recorded incidence, prevalence, and survival, of lung cancer for similar reasons, which should be explored by future research.

Comment 6:
6) More information about the demographics of the study population is needed. According to the methods section, both databases include patient-level information on demographics and lifestyle data. I encourage the authors include this information, both for a summary of the study population as well as for providing lung cancer rates stratified by demographics and lifestyle factors (these seem highly relevant, based on the risk factors described in the introduction and discussion).

Reply 6:
We have carried out additional analysis where we have stratified demographics, incidence and prevalence for CPRD GOLD by demographic region (England, Northern Ireland, Scotland and Wales) and have added this information into the supplement for the characterization and incidence analysis and have updated the results and discussion reflecting these new analyses.

As discussed in comment 3, smoking (as well as alcohol consumption) is poorly recorded in CPRD Aurum and GOLD. Furthermore, due to the time varying nature of smoking it is not possible to take this into account when estimating incidence rates without significant biases such as misclassification bias. However, we have stratified smoking status on the patient characteristics to compare those with a diagnosis of lung cancer and who are smokers, previous smokers, never smokers and missing to compare the demographics and hope this addresses the reviewers' comments.

Changes in the text:
New tables: Supplement S3 - S4
New Figures Supplement S7, S10, S13 .

Page 8, Lines 198-199: "Further stratifications of patient characteristics by UK region, and smoking status for CPRD GOLD can be found in Supplement S3 and S4."

Page 9 Lines 224-226
"show, after 2004, a decreasing trend in incidence for males, whereas an increase in incidence over the study period for females. Further stratification by UK region showed a similar trend for males and females (Supplement S7)"

Page 10 Lines 251-252
"Furthermore, stratification by UK region for GOLD showed similar trends across the different regions (Supplement S10)"

Page 11 Lines 278-279
"with similar results across the different UK regions (Supplement S13)"

Minor issues:
Comment 7:
• In the abstract, it seems somewhat irrelevant to mention the new screening program that was introduced in 2023 (page 1, line 14) because most of the study does not relate to this (see comments 4 and 5 above).

Reply 7: Thank you very much for your comment, which we understand and agree with. We have removed both references.

Changes in the text: Page 2, Deletion of: Introduction "as well as the introduction of new screening in 2023. AND Conclusion, Line 3. Deletion of "With the introduction of the UK lung cancer screening programme".

Comment 8:
• Are there differences between the two databases other than where the data were collected (UK for CPRD Gold, and England for Aurum)?

Reply 8: The main difference that the reviewer correctly points out is that the data have been collected for the whole of the UK for GOLD and only England for AURUM. Additionally, for GOLD and Aurum the GP practices use the Vision® or EMIS® software systems respectively. Due to some differences in structure and clinical coding in these two systems, these data are provided as two separate primary care databases. We have added further information to the manuscript to make this point clearer.

Changes in the text:
Page 5, Lines 124-129, add: from GP practices using the Vision® software system, whereas Aurum only contains data from England, from GP practices using the EMIS® software system… Uniquely, the mapping of both to a common data model (OMOP) allowed us to analyse both simultaneously and using the same exact analytical code.

Comment 9:
• Are there differences in study outcomes based on the database?

Reply 9: Thank you for your question. In this study we compare the study outcomes between both databases and show similar results. It is possible there could be some

differences in study outcomes between the different databases due to differences underlying source coding systems for GOLD, However, differences could also be related to the demographics of the different databases which we have addressed in an earlier comment from the reviewer (comment 6). Both databases have been mapped to a common data model, where GOLD has been mapped from Read codes to SNOMED codes (standard codes for the common data model we use), whereas AURUM coding is already using SNOMED codes for diagnosis codes.

We have adequately addressed the differences and similarities between the two data sources in the manuscript – full results are available in the supplementary, and throughout the manuscript we refer to both. This especially includes sections 3.2 (incidence), 3.3 (prevalence), and 3.4 (survival), where both are discussed in turn, where appropriate. In our discussion we discuss that the overall similarity is a strength. We have already added information about the differences between the databases based on reviewer comments.

References:

Common data model: https://www.ohdsi.org/data-standardization/

Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015 Jun 6;44(3):827–36. https://doi.org/10.1093/ije/dyv098

Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, Myles P. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol. 2019 Mar 11. pii: dyz034. https://doi.org/10.1093/ije/dyz034

Changes in the text: None.

Reviewer B

Comment 10: This is a well-written paper on trends in lung cancer incidence, survival, and prevalence in the United Kingdom. However, I have a major concern on results of this study: incidence rates are generally lower and trends are sometimes quite different from those reported by population-based cancer registries (for example, please see https://master-7rqtwti-hreqyzlibi4ac.uk-1.platformsh.site/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence#heading-Two). For example, the lung cancer incidence rate among males in the UK in 2016-2018 was about 90 per 100,000 according to cancer registry data, but according to Figure 1 of this manuscript, it was about 65 per 100,000; or lung cancer incidence rates among males have been declining according to cancer registry data, but in this study, rates have increased since 2002. I wonder if this is because the datasets used by the authors do not capture all cancer cases, or there are some issues with the corresponding population data. For this reason, I am not sure if the incidence data presented in this manuscript is informative.

Reply 10: In this manuscript we presented crude incidence rates to highlight the health care provision in the UK over time where we stratify by sex and by age. The statistics from the Cancer Research UK resource the reviewer cites have been age standardized to the European Standard Population 2013, which is why they differ from ours. However, we did compare the crude rates that are provided by CRUK and more recently with those estimated from the National Lung Cancer Audit 2019 and 2020 Rapid Cancer Registration Datasets (both cited in the manuscript). Our results are broadly in line with the latter (estimated from the national lung audit), however are lower compared to the CRUK statistics. This difference could be partly explained by misclassification, as the reviewer points out, in CPRD GOLD as we did not have linkage to cancer registry datasets. Furthermore, CRUK included cancers of the lung/bronchus and trachea whereas our study and the study using the lung cancer audit datasets did not include cancers of the trachea which also could contribute to the differences.

As we provided the trends by age group, we did not age standardize the incidence rates in the main manuscript. However, we realise this was not made clear in the methods and results so have added some text to highlight that we did carry out this analysis. Furthermore, the results from CRUK are taken from cancer registry data and therefore do not have a denominator population of individuals without a cancer diagnosis which is essential to calculate incidence. Their denominator population is instead taken from national population statistics which could introduce biases in the results and contribute to the differences between prior incidence rate calculations and those in the presented manuscript. We have stated this reasoning, with references, in the discussion of the manuscript. Additionally, a sharp increase at the beginning of the study period is likely caused by the introduction of the Quality and Outcomes Framework (QOF) which is a voluntary annual reward and incentive programme for all GP practices and promoted better recording of a variety of diseases including cancer.

We have updated the text to make these limitations clearer and have included more information regarding the age standardization.

References:

https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data

Changes to text:

Page 7, Lines 167-174.

"Age-standardized IRs were calculated using the 2013 European Standard Population (ESP2013) (REF). The ESP2013 serves as a standard population with a predefined age distribution which enables us to account for differences in age structures between different populations to ensure fair comparisons. The ESP2013 provides predefined age

distribution in five-year age bands; therefore, we collapsed these to obtain distributions for 10-year age bands used in this study. We used the age distribution of 20-29 years from ESP2013 for age-standardization as age distributions were not available for 18-29 years age band used in this study. "

Page 12 Lines 305-309
"Incidence rates reported here are broadly in line with those estimated from analyses from the Rapid Cancer Registration Databases however are lower than National Cancer Statistics (24-26). However, it must be noted that National Cancer Statistics include cancers of the lung/bronchus and trachea whereas this study and the Rapid Cancer Registration Databases only include cancers of the lung/bronchus."

Page 17, Lines 418-419.
"Firstly, we used primary care data without linkage to a cancer registry which could lead to misclassification, delayed recording, or incompleteness of cancer diagnoses(65) which could results in lower estimates"

Minor comments.

Comment 11:
The information provided in the Introduction appears to give the same weight to reduction in smoking prevalence and advances in treatment with regard to declines in lung cancer mortality. Although advances in treatment have been important, many of them have occurred more recently, and they are unlikely to explain a major part of declines in mortality from the 1980s to 2000s. Different trends in lung cancer mortality by sex, which largely reflects different historical patterns of smoking in males and females, provide further support for a much greater effect of changes in smoking prevalence on changes in lung cancer mortality; advances in treatment are expected to affect both sexes and is unlikely to explain different trends by sex.

Reply 11: Thank you very much for your comments. We believe that the introduction portrays a fair balance of smoking vs other factors in regard to smoking mortality and likely risk factors, which cannot be explained by smoking alone due to discussed and referenced discrepancies in smoking prevalence vs 10–20-year delayed mortality. Indeed, we state "Although smoking remains the greatest risk factor for all lung cancer subtypes".

Our study focusses on incidence/prevalence/survival from 2000-2021. We agree the most profound improvements have occurred more recently – but these are relevant to the survival curves in our study period. Indeed, the comparison we give is from the "mid-1980s to late-2010s" - which certainly includes surgical, medical, and multidisciplinary team-based changes in lung cancer treatment, as referenced in our manuscript.

We have added a sentence to the discussion stating that observed changed in age/sex over time could also contribute to observed differences in mortality

Changes to text:

Page 15 lines 378-380
"Furthermore, changes in age and sex distributions of those diagnosed with lung cancer over time could also contribute to the observed differences in survival."

Comment 12:
In the "Highlight Box", the authors state that "With the introduction of the UK lung cancer screening programme in 2023, this study will enable future comparisons of overall disease burden, so the impact of screening may be seen." However, the information provided in this manuscript is not enough to evaluate the impact of screening, and additional information (such as advances in treatment, behavioral factors, and so on) should be taken into account.

Reply 12: Thank you very much for this suggestion, we agree and have amended the highlights as shown below.
Changes to text:

What is known and what is new?
Lung cancer is the leading cause of cancer-associated mortality worldwide. In the UK, there has been a major reduction in smoking, the leading risk factor for lung cancer, alongside advances in lung cancer treatments, and changes in baseline population demographics.
With the introduction of the UK lung cancer screening programme in 2023, an up-to-date assessment of the trends of lung cancer before the introduction of this screening programme is required.

What is the implication, and what should change now?
~~**With the introduction of the UK lung cancer screening programme in 2023,**~~ **This study will help to enable future comparisons of overall disease burden, and changes in trends in the incidence, prevalence, and survival with lung cancer, so the impact of screening alongside novel treatments, and changing demographics may be seen.**
Further research is required to understand increasing disease burden in females.

Reviewer C

The manuscript is interesting, but statistical methods are only descriptive. I suggest major revisions.

Comment 13:

Line 82 should be in methods, not introduction.
Reply 13: Thank you for your comments. The TLCR Author Instructions state "A statement should be included at the end of the "Introduction" to indicate which reporting checklist was followed (e.g., "We present this article in accordance with the CONSORT reporting checklist.")**."** Therefore, we tink this should stay in the introduction.
Change to text: None.

Comment 14:
Line 129 to 135: please use standardized rates.
Reply 14:
We have addressed this point in a previous reviewer comment and have included age standardized results (point 10 reviewer B).
Change to text: See comment 10 reviewer B

There are several limitations, as partially discussed by the authors:
Comment 15:
Lack of Cancer Registry Linkage: Without direct linkage to cancer registries, there is a potential for misclassification or delayed recording of diagnoses, although the study asserts high accuracy in primary care records.

Reply 15: Thank you for your comment. The reviewer is correct, without direct linkage, there is the potential for misclassification and delay in diagnosis. However, a recent study using CPRD GOLD (reference below) showed that the positive predictive value of cancer particularly lung cancer is greater than 80% when using CPRD GOLD alone without linkage to cancer registry/ hospital data however a lower sensitivity. We have elaborated on this further in the discussion.

Change to text: Page 17 Line 419-422: Add "A previous validation study has shown high accuracy and completeness of cancer diagnoses in primary care records. However, although a high positive predictive value (>80%) was observed, a lower sensitivity was also reported (66) "

References
Strongman, Helen, Rachael Williams, and Krishnan Bhaskaran. "What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? a concordance and validation study using linked English electronic health records data." BMJ open 10.8 (2020): e037719.

Comment 16:
Exclusion of Detailed Cancer Characteristics: The absence of data on tumor histology, genetic mutations, staging, and specific treatments means that survival estimates might not fully account for variability in outcomes based on these factors.

Reply 16: Thank you for your comment, this is a valid point which we have already stated in the manuscript limitations. However, we have reworded this sentence to make it clearer.

Change to text:
Page 17 Lines 422-425
"Secondly, our use of primary care records, means we do not have information on tumor histology, genetic mutations, staging, and specific treatments, which means that survival estimates might not fully account for variability in outcomes based on these factors. Therefore, our survival estimates may overestimate survival in those with higher staging as well as those with specific subtypes or mutations associated with poorer survival such as SCLC (66)"

Comment 17:
Incomplete Smoking Status Data: With missing smoking status for a significant portion of the patients, there's an underrepresentation of this critical risk factor's role, although the reported prevalence among patients is notably high.

Reply 17: Thank you for your comment – we agree and updated this in the text.

Changes in the text: Page 19, Line 439-442 "Although not possible in CPRD due to missingness which could create biases in results, future research using alternative databases could be used to quantify the relative contribution of smoking in 'pack-years', alongside alternative discussed risk factors."

Comment 18:
You define this as a cohort study, but you should define it as a descriptive (cohort) study.

Reply 18: Thank you very much for your comment, we have changed the text as described below.
Change to text:
Abstract – Methods, Line 1. (overall line 54)
Methods – Section 2.1, Line 1. (overall line 116)

Comment 19:
You only describe incidence, prevalence, and survival across different groups without using regression models. You don't even use standardized rates or descriptive statistical tests. To fully capitalize on the potential of the data, statistical analysis should be updated applying proper statistical methods and possibly regression models.

Reply 19: As this is a descriptive study the aim was to describe the secular trends over time therefore regression models are out of the scope of this study. We have included age

standardized results in the supplement; however, we realize this was not clear in the methods or results so have updated the manuscript accordingly.

Change to text: Please see comment 10, Reviewer B which covers this comment.

Comment 20:
Results about gender are interesting, please better discuss it and the role of environmental pressures, educational level, and marital status (DOI: 10.3389/fpubh.2023.1278416, DOI: 10.4081/monaldi.2019.1017). I suggest expanding upon the discussion about the limitations of the study.

Reply 20:
Thank you for the additional suggestions and the reference. We have amended the manuscript the limitations in the discussion and included the suggested reference.

Changes to text: Page 16, Lines 421-425
However, although a high positive predictive value (>80%) was observed, a lower sensitivity was also reported(66). Secondly, our use of primary care records, means we do not have information on tumour histology, genetic mutations, staging, specific treatments, and further environmental factors, which means that survival estimates might not fully account for variability in outcomes based on these factors (67).

To conclude, this study is an interesting contribution to the epidemiological literature on lung cancer in the UK, offering comprehensive insights into the disease's dynamics over two decades. Despite its limitations, the research underscores the critical need for targeted public health interventions and continued improvement in lung cancer care and treatment.
Suggest: major revisions

We thank the reviewer for the positive assessment of our manuscript and hope the changes we have made are adequate and have improved the manuscript.

Reviewer D

Comment 21:
Abstract
line 23 - provide numbers for increasing incidence among females over 50 (compared to under 50), and males over 80 (vs. under).

Reply 21: We have provided numbers in the abstract as the reviewer has suggested. Although we were not able to directly compare the age groups suggested by the reviewer as we used 10 year age bands we compared these age bands and stated a range of values showing the increase in incidence over the study period.

Change to text:
Abstract
Females aged over 50 years of age showed increases in incidence over the study period, ranging from increases of 8 to 123 per 100,000 person-years, with the greatest increase in females aged 80-89. Alternatively, for males, only cohorts aged over 80 showed increases in incidence over the study period. The highest incidence rate was observed in people aged 80-89

Comment 22:
Introduction
line 65-67 - run-on sentence. Please edit to improve readability.

Reply 22: Thank you for your comment, we have updated the sentence to improve readbility as below.
Change to text: Page 4 Line 96. Delete "while", add full stop, to form:
This reduction is driven by a decrease in male mortality. Female mortality has slightly increased from 1970, peaking in 2010, despite the prevalence of smoking falling in both sexes (13).

Comment 23:
Methods:
Line 112 - were metastatic lung cancers also excluded by this reasoning?
Line 136-139 - repetitive information already mentioned on page 3.

Reply 23: Line 112 – thank you very much for your comment. Patients with Lung cancers that metastasised *to* the lung were excluded, but primary lung cancers that metastasised to *secondary organs* were included. We have updated the text to make this clearer.

Line 136-139 -
Change to text: Deleted both references to "≥ 18 years"
Page 6 Line 143: Change to "Diagnostic codes indicative of either non-malignant cancer or secondary metastases from other organs,"

Comment 24:
Is there any information on patient staging, method of resection, neoadjuvant or adjuvant therapy, number of lung cancers, pathology, lymph node involvement? This information would be useful to have in a study evaluating the prevalence of lung cancer and overall outcomes. If not, consider reducing the results section by 50% for brevity.

Reply 24: Thank you for your comments. As we discuss in our limitations, this information is not available in CPRD GOLD/Aurum. We think the included results in this manuscript are important to present the disease burden of lung cancer with age and sex stratifications. However, we have edited the results to reduce the length and remove repeating and

redundant sentences.
Change to text: Shortening of results section.

Comment 25:
Discussion:
Line 291 - It would be helpful to comment on method of treatment and staging to help the readership understand the factors that have contributed to improved survival (despite implementation of lung cancer screening only as of 2023).

Line 299 - the rise of lung cancer may attributed to a rise in the perceived incidence of lung cancer, such as improved imaging modalities and detection, screening, and awareness. It would be helpful to address these topics to inform the audience about the patterns of care in the UK.

Reply 25: Thank you very much for your comments. In response to previous reviewer comments we have expanded this paragraph on lines 381-393 which now addresses your helpful comment, which we agree with.

With regard to line 291 specifically, we have addressed this point already in the manuscript on lines 414-419.

Change to text: Page 15: The increases in survival, incidence, and prevalence, in this study could also be due to better diagnostic methods and improved public awareness of lung cancer symptoms, leading to earlier detection. For instance, a 2012 UK Department of Health campaign was implemented to raise awareness of persistent cough as a lung cancer symptom, leading to a 3.1% increase in the proportion of NSCLC diagnosed at stage one (58), with similar campaigns since (59). If diagnoses are indeed occurring at an earlier stage, lead-time bias may result in improvements in the survival data that do not exist in practice due to the fact the cancer was simply detected earlier, even if the treatment given and ultimate date of death is unchanged (60). Of course, it may be the case that whilst cancers are diagnosed sooner, this is coupled with better treatment, not only for an equivalent stage due to improved therapy but also because even better treatment may be delivered at the earlier stage of cancer. The introduction of CT screening may increase the recorded incidence, prevalence, and survival, of lung cancer for similar reasons, which should be explored by future research.

Comment 26:
Figure 2 - suggest including these figures in supplement and only include cumulative incidence in the manuscript.

Reply 26: Thank you very much for your comment. As we present crude incidence rates, the division into separate age groups provides necessary information in this context and therefore warrants them in the main manuscript not the supplement.

Change to text: None.