Original Article

Incidence, prevalence, and survival of lung cancer in the United Kingdom from 2000–2021: a population-based cohort study

George Corby[1], Nicola L. Barclay[1], Eng Hooi Tan[1], Edward Burn[1], Antonella Delmestri[1], Talita Duarte-Salles[2,3], Asieh Golozar[4,5], Wai Yi Man[1], Ilona Tietzova[6], Daniel Prieto-Alhambra[1,3]^, Danielle Newby[1]

[1]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Institute for Musculoskeletal Sciences, University of Oxford, Oxford, UK; [2]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; [3]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; [4]Odysseus Data Services, Cambridge, MA, USA; [5]OHDSI Center at the Roux Institute, Northeastern University, Boston, MA, USA; [6]First Department of Tuberculosis and Respiratory Diseases, First Faculty of Medicine, Charles University, Prague, Czech Republic

Contributions: (I) Conception and design: All authors; (II) Administrative support: A Delmestri, WY Man; (III) Provision of study materials or patients: D Prieto-Alhambra, A Golozar, I Tietzova; (IV) Collection and assembly of data: A Delmestri, WY Man; (V) Data analysis and interpretation: G Corby, E Burn, D Prieto-Alhambra, D Newby; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Daniel Prieto-Alhambra, PhD. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Institute for Musculoskeletal Sciences, University of Oxford, Windmill Road, Oxford OX3 7LD, UK; Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands.

Background: Lung cancer is the leading cause of cancer-associated mortality worldwide. In the United Kingdom (UK), there has been a major reduction in smoking, the leading risk factor for lung cancer. Therefore, an up-to-date assessment of the trends of lung cancer is required in the UK. This study aims to describe lung cancer burden and trends in terms of incidence, prevalence, and survival from 2000–2021, using two UK primary care databases.

Methods: We performed a population-based cohort study using the UK primary care Clinical Practice Research Datalink (CPRD) GOLD database, compared with CPRD Aurum. Participants aged 18+ years, with 1-year of prior data availability, were included. We estimated lung cancer incidence rates (IRs), period prevalence (PP), and survival at 1, 5 and 10 years after diagnosis using the Kaplan-Meier (KM) method.

Results: Overall, 11,388,117 participants, with 45,563 lung cancer cases were studied. The IR of lung cancer was 52.0 [95% confidence interval (CI): 51.5 to 52.5] per 100,000 person-years, with incidence increasing from 2000 to 2021. 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 years. Alternatively, for males, only cohorts aged over 80 years showed increases in incidence over the study period. The highest IR was observed in people aged 80–89 years. PP in 2021 was 0.18%, with the largest rise seen in participants aged over 60 years. Median survival post-diagnosis increased from 6.6 months in those diagnosed between 2000–2004 to 10.0 months between 2015–2019. Both short and long-term survival was higher in younger cohorts, with 82.7% 1-year survival in those aged 18–29 years, versus 24.2% in the age 90+ years cohort. Throughout the study period, survival was longer in females, with a larger increase in survival over time than in males.

Conclusions: The incidence and prevalence of lung cancer diagnoses in the UK have increased, especially in female and older populations, with a small increase in median survival. This study will enable future comparisons of overall disease burden, so the overall impact may be seen.

Keywords: Lung cancer; incidence; prevalence; cancer survival

^ ORCID: 0000-0002-3950-6346.

Submitted Mar 18, 2024. Accepted for publication Jul 03, 2024. Published online Sep 21, 2024.
doi: 10.21037/tlcr-24-241
View this article at: https://dx.doi.org/10.21037/tlcr-24-241

Introduction

Lung cancer is the leading cause of cancer-associated mortality worldwide, with 1.8 million deaths in 2020 (1). New diagnoses are predicted to nearly double by 2070 meaning it will continue to be a major cause of morbidity and mortality globally (2).

As early as 1962, the British Royal College of Physicians (RCP) published *Smoking and Health* (3), which established a clear and important link between smoking and lung cancer. Although smoking remains the greatest risk factor for all lung cancer subtypes, other factors such as family history (4), and exposure to arsenic (5), radon (6), biomass fuels, asbestos (7), or a broad range of occupational chemicals, have been identified as increasing risk. These risk factors can influence the proportion of patients with each subtype of lung cancer (2).

The United Kingdom (UK) Office of National Statistics (ONS) surveys indicate that 60 years from the initial RCP report, smoking prevalence in the UK continues to decrease, from 45% in the early 1970s to 14% in 2020 (8), with a concurrent rise in the use of electronic cigarettes (9,10). This reduction in the prevalence of smoking has coincided with a revolution in treatments of lung cancer, with major advancements in surgery (11), targeted drug treatments (12), and a shift towards multidisciplinary team (MDT) management (13). Together, this has corresponded with a major fall in mortality rates of lung cancer, with a 38% fall in mortality rate per 100,000 from the mid-1980s to late-2010s (13). 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).

In 2023, the UK introduced targeted screening for lung cancer. People aged 55–74 years with a general practice (GP) health record documenting a smoking history will be invited to interview for a risk assessment, after which they may be offered a low-dose computed tomography (CT) scan. When the rollout is complete, it is estimated 325,000 people will become newly eligible for a scan every year. Focussing on those with the highest risk will enable earlier diagnosis and potentially better survival, with reduced iatrogenic harm from screening-related radiation exposure (14,15).

Due to changes in risk factor exposure, and the introduction of the new screening in 2023, a comprehensive assessment of the trends of lung cancer in different population strata using routinely collected data from primary care, is required in the UK. Understanding these trends in lung cancer is an important aspect of population healthcare planning, particularly considering the introduction of screening high-risk individuals. Therefore, the aim of this study is to describe lung cancer burden and trends in terms of incidence, prevalence, and survival from 2000–2021 using two large and representative primary care databases from the UK. We present this article in accordance with the STROBE reporting checklist (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-241/rc).

Highlight box

Key findings

- The incidence and prevalence of lung cancer diagnoses in the United Kingdom (UK) have increased, especially in female and older populations, with a small increase in median survival. Both short- and long-term lung cancer survival has improved over time for both sexes.

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?

- 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.

Methods

Study design, setting, and data sources

We carried out a population descriptive cohort study using routinely collected primary care data from the UK. People

with a diagnosis of lung cancer and a background cohort (denominator population) were identified from Clinical Practice Research Datalink (CPRD) GOLD to estimate overall survival, incidence, and prevalence. We additionally carried out this study using CPRD Aurum to compare the results for GOLD. Both these databases contain pseudonymised patient-level information on demographics, lifestyle data, clinical diagnoses, prescriptions, and preventive care provided to patients and collected by the National Health Service (NHS) as part of their care and support. CPRD GOLD contains data from across the UK, from GP practices using the Vision® software system, whereas Aurum only contains data from England, from GP practices using the EMIS® software system. Both databases are established primary care databases covering over 50 million people (16), and both were mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) (17,18). Uniquely, the mapping of both to a common data model (OMOP) allowed us to analyse both simultaneously, and using the same exact analytical code. The use of CPRD data in this study was approved by the Independent Scientific Advisory Committee (No. 22_001843). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Individual consent for this retrospective analysis was waived.

Study participants and time at risk

All participants were required to be aged 18 years or older and have at least 1-year of prior history. For the incidence, prevalence, and survival analyses, the study cohort consisted of individuals present in the database from 1st January 2000. For CPRD GOLD, these individuals were followed up to whichever came first: practice stopped contributing to the database, patient left the practice, date of death, or the 31st of December 2021 (the end of the study period) whereas for Aurum, the end of the study period was 31st of December 2019. For the survival analysis, only individuals with newly diagnosed lung cancer were included. Any patients whose death and cancer diagnosis occurred on the same date were removed from the survival analysis.

Lung cancer definition

We used Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) diagnostic codes to identify lung cancer events. Diagnostic codes indicative of either non-malignant cancer or secondary metastases from other organs, were excluded as well as diagnosis codes indicative of melanoma and lymphoma occurring in the organs of interest. The study outcome cancer definition was reviewed with the aid of the CohortDiagnostics R package (19). This package was used to identify additional codes of interest and to remove those highlighted as irrelevant based on feedback from clinicians with oncology, primary care, and real-world data expertise through an iterative process during the initial stages of analyses. The clinical code lists used to define lung cancer can be found in Table S1.

OMOP-based computable phenotypes are available, together with all analytical code at our GitHub repository to enable reproducibility (see statistical methods). For survival analyses, mortality was defined as all-cause mortality based on CPRD GOLD date of death records, which have been previously validated and shown to be over 98% accurate (20).

Statistical analysis

The population characteristics of patients with a diagnosis of lung cancer were summarised, with median and interquartile range (IQR) used for continuous variables and counts and percentages used for categorical variables.

We calculated the overall and annualised crude incidence rates (IRs) and annualised prevalence for lung cancer from 2000 to 2021. For incidence, the number of events, the observed time at risk, and the IR per 100,000 person-years were summarised along with 95% confidence intervals (95% CI). Annual IR were calculated as the number of incident lung cancer cases as the numerator and the number of person-years in the general population within that year as the denominator whereas overall incidence was calculated from 2000 to 2021.

Age-standardized IRs were calculated using the 2013 European Standard Population (ESP2013) (21). The ESP2013 serves as a standard population with a predefined age distribution where to account for differences in age structures between different populations to ensure fair comparisons. The ESP2013 provides predefined age distribution in 5-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.

Period prevalence (PP) was calculated on 1st January for the years 2000 to 2021, with the number of patients fulfilling the case definition for lung cancer as the

numerator. The denominator was the number of patients on 1st January in the respective years for each database. The number of events, and prevalence (%) were summarised along with 95% CIs.

For survival analysis, we used the Kaplan-Meier (KM) method to estimate the overall survival probability from observed survival times with 95% CIs. We estimated the median survival and survival probability 1, 5, and 10 years after diagnosis.

All results were stratified by database, by age (10-year age bands apart from the first and last age bands which were 18 to 29 years and 90 years and older respectively) and by sex. Additionally, for GOLD only, we stratified by calendar time of cancer diagnosis (2000–2004, 2005–2009, 2010–2014, 2015–2019 and 2020–2021) to understand if survival has changed over time. To avoid re-identification, we do not report results with fewer than five cases.

For Aurum, the same statistical analyses were performed using data from 1st January 2000 to 31st December 2019 to compare the results obtained from GOLD.

The statistical software R version 4.2.3 was used for analyses. For calculating incidence and prevalence, we used the IncidencePrevalence R package (22). For survival analysis, we used the survival R package (23). All analytic code used to perform the study is available at https://github.com/oxford-pharmacoepi/EHDENCancerIncidencePrevalence.

Results

Patient populations and characteristics

Overall, there were 11,388,117 eligible patients identified from January 2000 to December 2021 from CPRD GOLD. Attrition tables can be found in the Table S2. A summary of baseline patient characteristics of those with a diagnosis of lung cancer is shown in *Table 1*. Further stratifications of patient characteristics by UK region, and smoking status for CPRD GOLD can be found in Tables S3,S4.

Overall, from the 45,563 patients with a diagnosis of lung cancer, patients were more likely to be male (54%), with a median age of 72 years old. The highest percentage of patients were aged 70–79 years old, contributing to 36.7% of diagnosed patients. Patients with lung cancer had a high prevalence of chronic obstructive pulmonary disease (COPD) (25%) as well as cardiovascular comorbidities such as heart disease (23.5%) and hypertensive disorder (27.2%) at presentation. Similar observations were seen across both databases. A similar table with detailed baseline characteristics for Aurum patients is available in Table S5.

For GOLD, stratification by UK region showed characteristics were generally similar across the different regions apart from smoking status which had lower proportions for England compared to the other regions (Table S3). Stratification on smoking status showed larger proportions of younger patients (40–59 years) who were smokers compared to non and former smokers, with the average age at presentation being 8 years older in non-smokers than smokers, 77 versus 69 respectively (Table S4). Smokers had higher proportions of diagnoses of COPD, depressive disorders, and peripheral vascular disease compared with the non-smoking group.

IRs stratified by calendar year, age, and sex

The overall IR of lung cancer from 2000 to 2021 was 52.0 per 100,000 person-years (95% CI: 51.5 to 52.5) for GOLD. Females had lower overall IR [47.2 per 100,000 person-years (95% CI: 46.5 to 47.8)] compared to males [57.0 per 100,000 person-years (95% CI: 56.3 to 57.7)], with similar rates in Aurum. Annualised IRs increased across both databases (*Figure 1*). Females showed increasing IRs over the study period, while males showed a more stable trend in both databases. Age standardized results using the ESP2013 for CPRD GOLD show, after 2004, a decreasing trend in incidence for males, whereas an increase in incidence over the study period for females (Figure S1). Further stratification by UK region showed a similar trend for males and females (Figure S2). All study results can be found in and downloaded from an interactive web application: https://dpa-pde-oxford.shinyapps.io/LungCancerIncPrevSurvShiny/.

Overall IRs increased with age up to 80–89 years across both databases (Table S6). Annualised IRs for each age group (*Figure 2*) over time showed for those aged 70–89 years a gradual increase between 2004–2019. Whereas for those 50–59 years of age there was a gradual decline in IRs since 2004. For those aged 60–69 years, IRs were stable from 2004–2019 for GOLD, whereas for Aurum, IRs increased from 2000 to 2013 before stabilising in 2019. For those over 90 years of age, there was also a stabilisation of IRs from 2014 with larger differences between the databases. Younger patients (30–49 years of age) showed relatively stable IRs over the study period.

Stratification by age and sex (Figure S3) showed similar trends across both databases. For females aged 60–89 years, IRs increased between 2000–2019, whereas for males, IRs

Table 1 Baseline characteristics of lung cancer patients at the time of diagnosis for CPRD GOLD

Database	CPRD GOLD
Number of lung cancer patients	45,563
Sex: male, N (%)	24,569 (53.9)
Age (years), median [IQR]	72 [65 to 79]
Age groups (years), N (%)	
18–29	24 (0.1)
30–39	129 (0.3)
40–49	1,030 (2.3)
50–59	4,660 (10.2)
60–69	12,249 (26.9)
70–79	16,745 (36.8)
80–89	9,546 (21.0)
90+	1,180 (2.6)
Prior history (days), median [IQR]	3,660 [2,009 to 5,352]
Smoking status (any time 5 years prior), N (%)	
Non-smoker	7,154 (15.7)
Former smoker	543 (1.2)
Current smoker	22,019 (48.3)
Missing	15,847 (34.8)
General conditions (any time prior), N (%)	
Atrial fibrillation	3,207 (7.0)
Cerebrovascular disease	3,840 (8.4)
Chronic liver disease	247 (0.5)
Chronic obstructive lung disease	11,163 (24.5)
Coronary arteriosclerosis	677 (1.5)
Crohn's disease	154 (0.3)
Dementia	772 (1.7)
Depressive disorder	6,397 (14.0)
Diabetes mellitus	5,321 (11.7)
Gastroesophageal reflux disease	1,261 (2.8)
Gastrointestinal haemorrhage	3,113 (6.8)
Heart disease	10,704 (23.5)
Heart failure	2,006 (4.4)
HIV	23 (0.1)
Hyperlipidemia	4,586 (10.1)
Hypertensive disorder	12,404 (27.2)
Ischemic heart disease	6,237 (13.7)
Obesity	928 (2.0)
Osteoarthritis	9,841 (21.6)
Peripheral vascular disease	2,260 (5.0)
Pneumonia	2,549 (5.6)
Pulmonary embolism	804 (1.8)
Renal impairment	6,156 (13.5)
Ulcerative colitis	181 (0.4)
Venous thrombosis	2,331 (5.1)

CPRD, Clinical Practice Research Datalink; IQR, interquartile range; HIV, human immunodeficiency virus.

were relatively stable across this period. Males had higher IRs compared with females apart from those 30–59 years of age, where IRs were similar between sexes. Also, similar IRs were seen for females and males aged 60–69 years from 2015 onwards for GOLD.

PP for study population with database, age, and sex stratifications

For the whole population, the PP for lung cancer in 2021 was 0.18% (95% CI: 0.17% to 0.18%) for GOLD. PP in 2019 was similar across both databases (~0.17%). Sex stratification showed PP in 2019 were slightly higher in females [0.175% (95% CI: 0.168% to 0.181%)] compared to males [0.156% (95% CI: 0.150% to 0.162%)] in GOLD with smaller differences in Aurum. In GOLD, PP has increased 2.7-fold from 2000 to 2021, with females having a larger fold increase (3.7-fold) in PP across the study period compared to males (2.0-fold), with similar observations seen in Aurum (*Figure 3*). Furthermore, stratification by UK region for GOLD showed similar trends across the different regions (Figure S4).

When stratifying by age group, PP in 2019 was highest in those 80–89 years of age (0.80% Aurum, 0.72% GOLD) with this age group seeing the largest change in PP over the study period (*Figure 4*). Overall, most age groups showed increases in PP over the study period for both databases.

Stratification by sex and age showed similar trends between males and females across age groups (Figure S5). For those aged 50–59 years of age, PP for males remained relatively stable over the study period from 2005, whereas PP for females increased. For those 60–69 years of age, PP

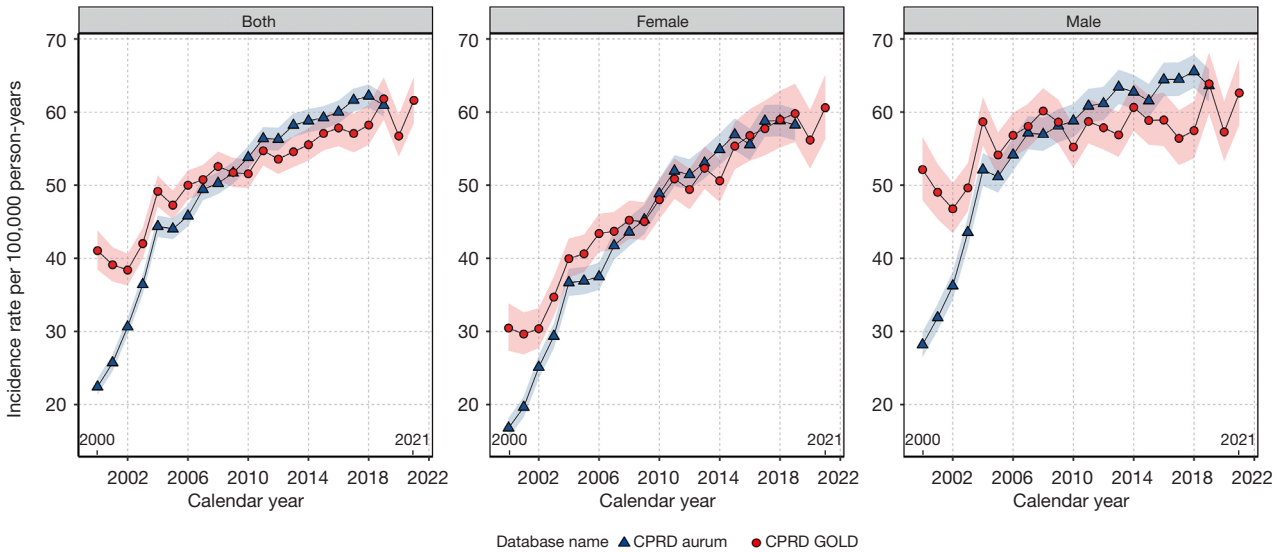

Figure 1 Annualised incidence rates for lung cancer from 2000 to 2021 stratified by database and sex. Bands show 95% confidence interval. CPRD, Clinical Practice Research Datalink.

increased for females from 2010, whereas for males in this age group PP initially increased from 2010 to then decrease from 2014 in GOLD (in Aurum PP continued to increase, albeit at a slower rate than in females). Overall, males had higher PP compared to females for those aged 80 years and older. For those aged 50–79 years males had higher PP earlier in the study period with females having higher PP towards the end of the study period. For those aged 40–59 years, there were no differences in PP over time between males and females.

Overall survival rates for the cancer population with age, sex, and calendar year stratification

For GOLD there were 43,903 patients with 35,381 deaths (80.6% of patients) over the study period with a median follow-up of 0.54 years (IQR, 0.18–1.39 years). For Aurum, there were 86,710 patients with 67,421 deaths (77.8% of patients) over the study period with a median follow-up of 0.58 years (IQR, 0.19–1.49 years).

Figure S6 shows survival curves for the overall population and stratified by sex. The median survival for the whole population was 0.66 years (95% CI: 0.64–0.67) and 0.71 years (95% CI: 0.70–0.72) in GOLD and Aurum respectively. Survival after 1, 5 and 10 years after diagnosis was 39.0%, 12.0% and 6.4% for GOLD, with similar results across the different UK regions (Figure S7) and similar values for Aurum.

Median survival was higher in females (0.75–0.81 years) compared to males (0.60–0.64 years) across both databases. Regarding short- and long-term survival, females had better survival compared with males (Table S7). When stratifying by age group, median survival decreased with age for both databases from 1.3–1.5 years for those aged 30–39 years to ~0.35 years for those aged 90 years and older (Table S8). Survival at 1, 5, and 10 years also decreased with increasing age (Table S9).

Median survival increased from 6.6 months in those diagnosed between 2000 to 2004 to 10 months for those diagnosed between 2015 to 2019 for both sexes (*Figure 5*, Table S10). Median survival was similar for those diagnosed in 2020–2021 to 2015–2019. For different age groups, median survival increased over time for most age groups (40–89 years of age) apart from those 90 years and older where median survival has not changed over time.

Overall, 1-year survival has increased between 2000–2004 and 2015–2019 for the whole population with similar trends for both sexes for GOLD. For the whole population, 1-year survival increased from 33% in 2000–2004 to 45% in 2015–2019. For 5-year survival, there were linear increases in survival over time for GOLD from 2000–2004 to 2015–2019 (Table S11). When stratifying by age group, both 1- and 5-year survivals were higher in those diagnosed between 2015–2019 compared to those diagnosed between 2000–2004 for those 50–89 years of age with similar patterns for both sexes.

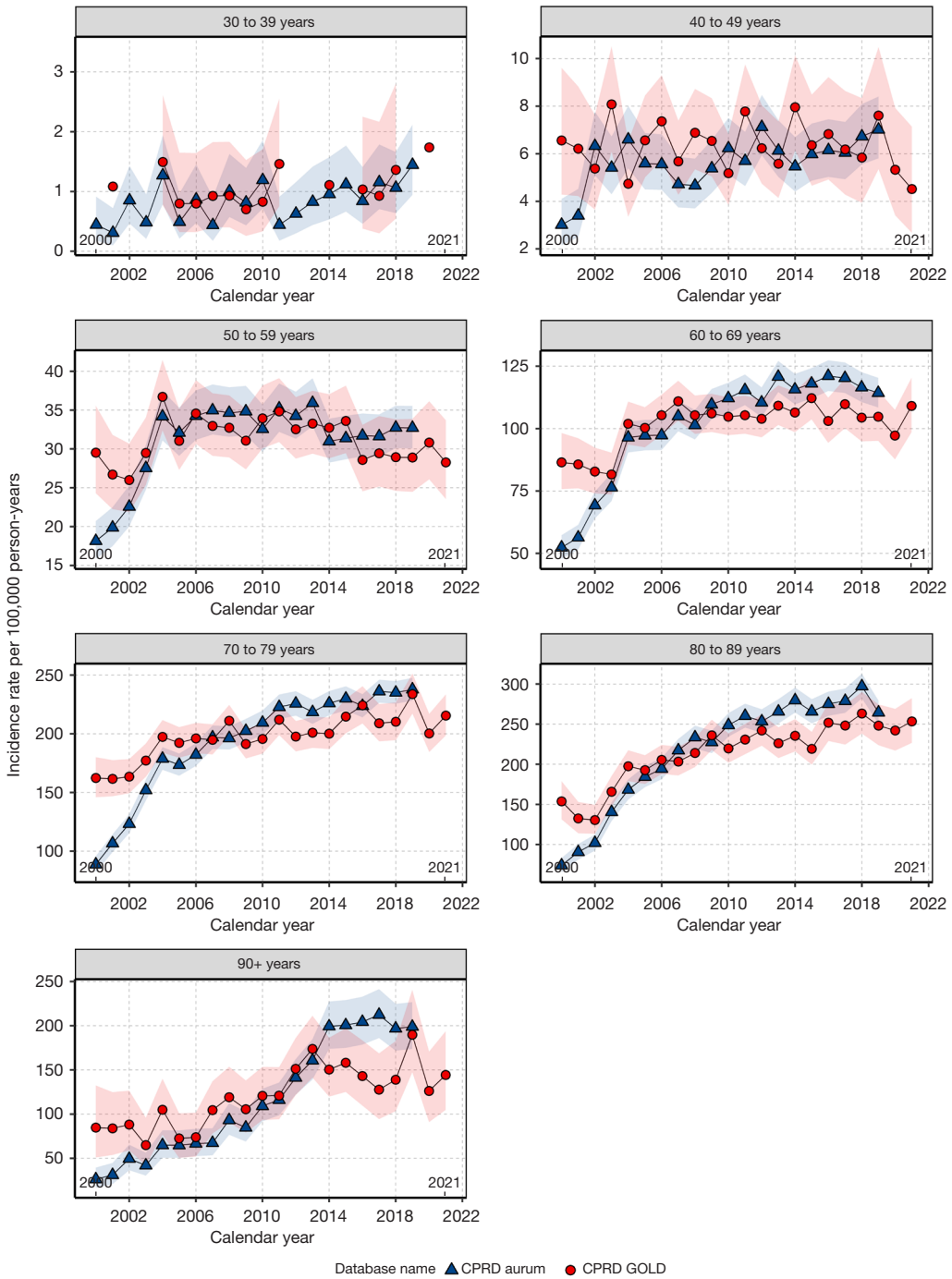

Figure 2 Annualised incidence rates from 2000 to 2021 stratified by database and age group. Bands show 95% confidence interval. CPRD, Clinical Practice Research Datalink.

Discussion

This study provides a comprehensive analysis of trends in lung cancer incidence, prevalence, and survival in the UK. In this large cohort of over 11 million people, the incidence and prevalence of lung cancers in the UK has increased from 2000 to 2021, while both short- and long-term survival has slightly improved across all age groups over this

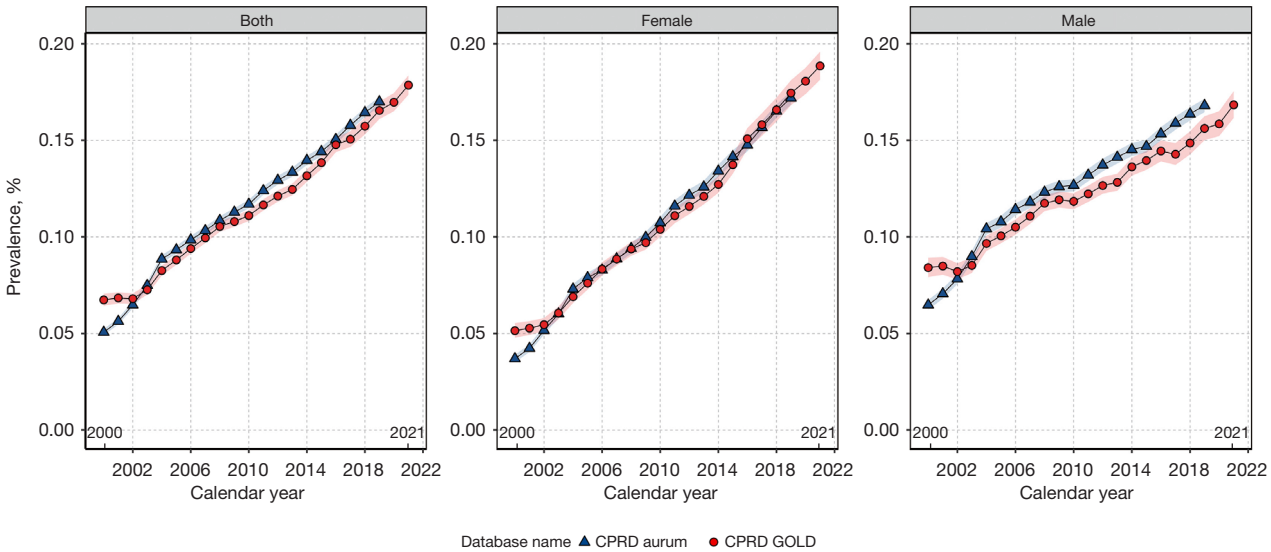

Figure 3 Annual period prevalence from 2000 to 2021 for the whole population and stratified by sex. Bands show 95% confidence interval. CPRD, Clinical Practice Research Datalink.

time period, with better survival in females than males.

IRs 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. For age and sex, studies have reported higher IRs in males compared to females and higher incidence with increasing age peaking in those aged 80 to 89 years of age in line with our estimates (26,27). Other studies have also shown increases in prevalence, particularly in females, where prevalence has overtaken males in recent years (1,28).

The increase in incidence and prevalence of lung cancer over time has been reported particularly in countries with higher levels of economic development as well as countries with higher smoking prevalence and air pollution (1,2,29). These increases have largely been driven by females. However, other studies show a decline or stabilisation of the incidence of lung cancer due to decreases in males (1,30,31). Our age standardized IRs to the European Standard Population 2013 are in line with these observations for females and males. For this work we report crude IRs trends overall and stratified by sex and age to understand health provision in the UK. IRs are known to vary by age as shown by this work and this is concordant with national cancer statistics (25) which show decreases in incidence over time in those aged 50–59 and aged over 80 years particularly after 2014 with increases in those aged 60–79 years. These age differences in time trends are likely due to birth cohort effects as well as differences in modifiable risk factor exposures (25). However, other studies also show IRs have decreased over time in all age groups which could be driven by a larger decrease in male incidence over time even with the increasing incidence in females (2,32).

Encouragingly, despite the increases in incidence and prevalence, our study shows 1- and 5-year survival has improved over the past 20 years, with increases of 11% and 7% respectively. Our results are in line with the 2023 National Lung Cancer Audit for England and Wales (33), UK cancer registry (34), as well as other international studies (1,35-37). However, despite these improvements, short- and long-term survival is still low compared with other common cancers such as breast and prostate. Interestingly, 5-year survival rates in this study are broadly comparable across Europe with values between 10–20% (1,35-37).

The rise in lung cancer could be for numerous reasons, making future trends difficult to predict. Smoking is the greatest risk factor for the disease, with a 20-year delay between smoke exposure and cancer onset (38). In our study, a high proportion of lung cancer patients (74%) were either current or ex-smokers; which is approximately five times higher than the overall population smoking prevalence (8) but slightly lower than UK national audit data estimates of

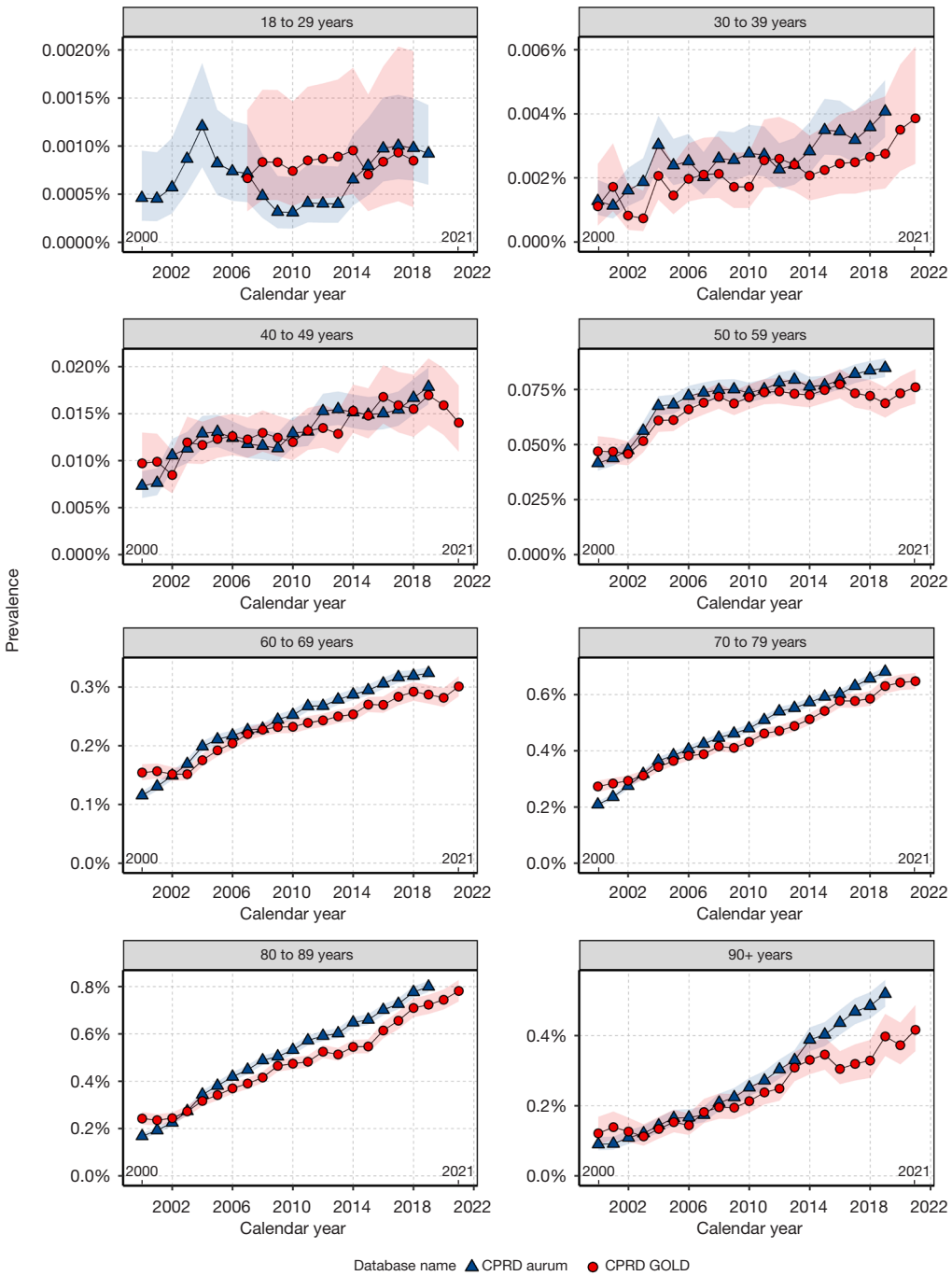

Figure 4 Annual prevalence from 2000 to 2021 stratified by database and age group. Bands show 95% confidence interval. CPRD, Clinical Practice Research Datalink.

smoking status in people with lung cancer (33). However, following numerous successful public health interventions, the prevalence of smoking has fallen in the UK (8) as well as many other countries across Europe (39), although may have increased during recent lockdowns (40). Furthermore, UK legislation has reduced the impact of 'second-hand' smoking which has amplified the impact of the reduction in smoking prevalence further in the UK (41). Therefore,

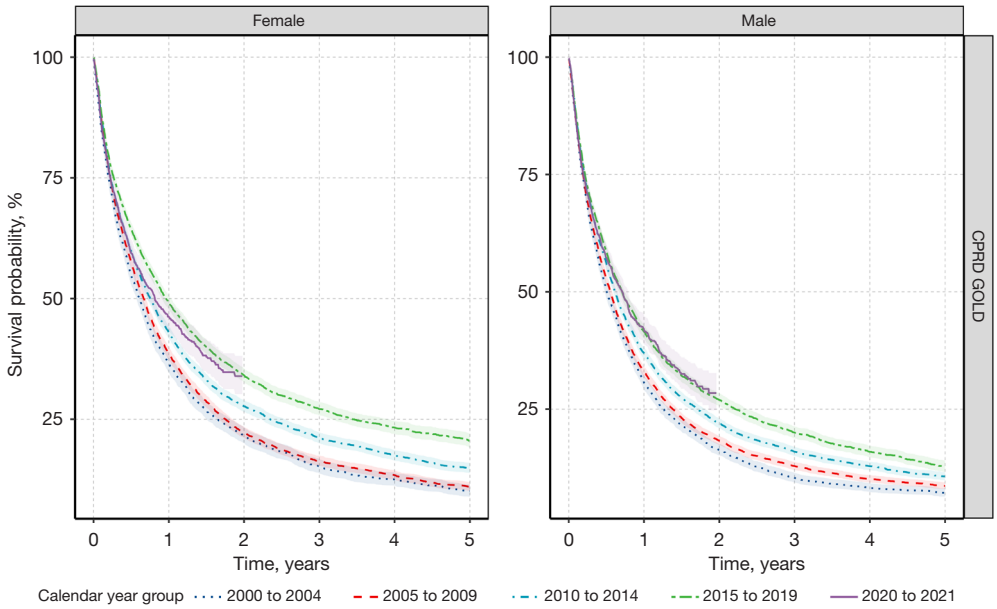

Figure 5 Kaplan-Meier survival curve of lung cancer stratified by sex and calendar year of diagnosis (2000–2004, 2005–2009, 2010–2014, 2015–2019 and 2020–2021). Bands show 95% confidence interval. CPRD, Clinical Practice Research Datalink.

increases in disease burden in this study cannot be attributed to smoking prevalence alone. However, in our study, we further see an overall shift, from 2001–2021, in the age distribution of incident diagnoses, from younger to older participants—with incidence rising especially in females aged over 50 years, and males over 80 years. This may reflect that lower exposure to cigarette smoke is causing people that are predisposed to developing lung cancer, to develop it at an older age.

Alternatively, in recent years, electronic cigarettes have been promoted as a harm reduction strategy to a greater extent than in many other countries. The effects of this are largely unknown, especially with the lag time before cancer onset, so the effect of this may not yet be full seen (42).

Other risk factors include COPD, which is an independent risk factor for lung cancer (2). Nearly a quarter (24.5%) of patients had a prior diagnosis of COPD across both sexes, which is higher than existing estimates of COPD prevalence in the overall population (43).

In never-smokers, defined as individuals who have smoked fewer than 100 cigarettes in their lifetime (39), radon is the leading environmental cause of lung cancer (44). As with smoking, there is a substantial exposure-cancer-mortality delay (45), and public health campaigns have aimed to reduce exposure (6). Air pollution is also an established risk factor for lung cancer (46) with a study using UK Biobank demonstrating additive interactions between air pollution and genetic-related risk factors for lung cancer (47). Despite apparent reductions of exposure to the discussed major modifiable risk factors in the UK, incidence and prevalence of lung cancer has increased in our study.

While not a sex-specific cancer, lung cancer shows sex-specific trends in our data as well as has been reported nationally and internationally (48-50). A recent review attributed this finding to many factors such as the slower relative reduction in female smoking compared to males (8,49), coupled with females being exposed to different risk factors, as well as differences in oestrogen levels (49), exposure to human papillomavirus (HPV) (51) and genetic polymorphisms (49). Furthermore, females have been reported to have a higher frequency of mutations in critical genes, such as TP53 and the KRAS leading to higher risk of the disease (52-54).

Although incidence and prevalence are increasing in females, overall survival is better in females compared to males in line with a similar study using primary care data from the UK (27). Again, reasons for this difference are likely to be multifactorial involving different risk factors, treatment decisions, and cancer histology (48). Furthermore, changes in age and sex distributions of those diagnosed with lung cancer over time could also contribute to the observed differences in survival. Survival from lung cancer has

substantially improved since 2000, which has corresponded with major advancements in lung cancer treatments (2,12). Towards the start of this period, management of this disease progressively included a multi-disciplinary team (MDT) for cancer care (11,55,56). Additionally, lung cancer treatments have further shifted towards more targeted regimens, after demonstrating better survival than existing therapy in trials (12). Additionally, there has been a concurrent expansion of lung cancer surgery, which has been shown to further improve treatment outcomes (57,58).

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 non-small cell lung cancer (NSCLC) diagnosed at stage 1 (59), with similar campaigns since (60). 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 (61). 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.

The main strength of this study is the use of a large primary care database covering the whole of the UK and validation of the results using another database from England. CPRD GOLD covers primary care practices from England, Wales, Scotland, and Northern Ireland whereas CPRD Aurum covers primary care practices in England. The similarity between the results in both databases, and their overall agreement with National Cancer Statistics (NCS) and national audit programmes provide increased generalizability across the UK. The sharp increase in incidence at the start of the study period between 2000–2004 could be due and the institution of cancer quality improvement measures by the NHS in 2003 as well as the introduction of the Quality and Outcomes Framework (QOF) in 2004 which encourages general practitioners to record all new cases of cancer which could partly explain the increase in cancer recording (62).

Another strength of our study is the inclusion of a complete study population database for the assessment of incidence and prevalence. In contrast, cancer registry studies extrapolate the registry data to the whole population using national population statistics, potentially introducing biases (63,64). The high validity and completeness of mortality data with over 98% accuracy compared to national mortality records (20) allowed us to examine the impact of calendar time on overall survival—one of the key outcomes in cancer care.

Our study also has some limitations. 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. 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). 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). 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 small cell lung cancer (SCLC) (68). Other factors, such as socio-economic status, environmental exposures such as air pollution and ethnicity could also result in different values for incidence, prevalence, and survival (46,47). Thirdly, in this study, we calculated overall survival, which does not differentiate between deaths caused by cancer *vs.* other causes. Therefore, it is a broad measure of overall survival rather than specifically cancer mortality. However, with the introduction of low-dose CT screening being introduced in the UK for targeted groups, the use of overall survival will enable the assessment of how much screening promotes overall survival and prevents all-cause mortality (including for instance changes related to non-lung cancer incidental diagnoses on CT screening), not just mortality related to lung cancer, and the overall benefit (or harm) associated with lung cancer treatments. Finally, smoking status was missing in 34.8% of lung cancer patients in this study with 48.3% of patients recorded as smokers which is around 3–5× the overall population prevalence of smoking for this period, however may be higher (8). 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.

Conclusions

Despite the falling prevalence of smoking in the UK, the incidence and prevalence of lung cancer is increasing. Reassuringly, the improvements in survival over the study period highlight the development of better-targeted treatments and earlier diagnosis of high-risk populations. However, the rise of lung cancer even with a decrease in smoking is a cause for concern, particularly in females. Further work needs to focus on understanding this demographic shift, to explain why lung cancer continues to rise, which could lead to better prevention, earlier diagnosis, and further targeted treatments. As the UK Lung Cancer Screening Programme is introduced, with a target of 100% implementation by 2030, our study has the potential to enable subsequent comparisons of not only survival rates, but the baseline medical characteristics of people at diagnosis ultimately enabling a more comprehensive assessment of the impact of the screening programme and resulting public health interventions in the UK.

Acknowledgments

Funding: This activity under the European Health Data & Evidence Network (EHDEN) and OPTIMA has received funding from the Innovative Medicines Initiative 2 (IMI2) Joint Undertaking under grant agreement No. 806968 and No. 101034347 respectively. IMI2 receives support from the European Union's Horizon 2020 research and innovation programme and European Federation of Pharmaceutical Industries and Associations (EFPIA). The sponsors of the study did not have any involvement in the writing of the manuscript or the decision to submit it for publication. Additionally, there was partial support from the Oxford NIHR Biomedical Research Centre.

Footnote

Reporting Checklist: The authors have completed the STROBE reporting checklist. Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-241/rc

Data Sharing Statement: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-241/dss

Peer Review File: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-241/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-24-241/coif). N.L.B. receives grants/consultancy fees/payments from Theramex, F. Hoffmann-La Roche, Innovate UK, and Sleep Universal Limited. A.G. is the Vice President, and Global Head of Data Sciences, at Odysseus Data Services Inc, since September 2021. D.P.A. research group has received research grants from Amgen, Chiesi-Taylor, Lilly, Janssen, Novartis, and from UCB Biopharma; consultancy fees (paid to his department) from Astra Zeneca and UCB Biopharma; and other financial or non-financial interests from Amgen, Astellas, Janssen, Synapse Management Partners and UCB Biopharma have funded or supported training programmes organised by D.P.A. and his department. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The use of CPRD data in this study was approved by the Independent Scientific Advisory Committee (No. 22_001843). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Individual consent for this retrospective analysis was waived.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.

2. Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. Nat Rev Clin Oncol 2023;20:624-39.
3. James J. Smoking, information, and education: The Royal College of Physicians and the new public health movement. Journal of Policy Analysis and Management 2024;43:446-71.
4. Ang L, Chan CPY, Yau WP, et al. Association between family history of lung cancer and lung cancer risk: a systematic review and meta-analysis. Lung Cancer 2020;148:129-37.
5. Palma-Lara I, Martínez-Castillo M, Quintana-Pérez JC, et al. Arsenic exposure: A public health problem leading to several cancers. Regul Toxicol Pharmacol 2020;110:104539.
6. Lorenzo-González M, Torres-Durán M, Barbosa-Lorenzo R, et al. Radon exposure: a major cause of lung cancer. Expert Rev Respir Med 2019;13:839-50.
7. Kwak K, Kang D, Paek D. Environmental exposure to asbestos and the risk of lung cancer: a systematic review and meta-analysis. Occup Environ Med 2022;79:207-14.
8. Office for National Statistics [Internet]. 2021. Smoking prevalence in the UK and the impact of data collection changes: 2020. Accessed 26 Feb 2024. Available online: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/drugusealcoholandsmoking/bulletins/smokingprevalenceintheukandtheimpactofdatacollectionchanges/2020
9. Martins BNFL, Normando AGC, Rodrigues-Fernandes CI, et al. Global frequency and epidemiological profile of electronic cigarette users: a systematic review. Oral Surg Oral Med Oral Pathol Oral Radiol 2022;134:548-61.
10. National Health Service, 2022. Accessed 26 Feb 2024. Available online: https://digital.nhs.uk/news/2022/decrease-in-smoking-and-drug-use-among-school-children-but-increase-in-vaping-new-report-shows#:~:text=The%20number%20of%20young%20people,1%2C%20-statistics%20published%20today%20show
11. Kowalczyk A, Jassem J. Multidisciplinary team care in advanced lung cancer. Transl Lung Cancer Res 2020;9:1690-8.
12. Michelotti A, de Scordilli M, Bertoli E, et al. NSCLC as the Paradigm of Precision Medicine at Its Finest: The Rise of New Druggable Molecular Targets for Advanced Disease. Int J Mol Sci 2022;23:6748.
13. Cancer Research UK, [Internet]. 2024. Lung cancer mortality statistics. Accessed 26 Feb 2024. Available online: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/mortality#heading-Two
14. UK Government, June 2023, Accessed 26 Feb 2024. Available online: https://www.gov.uk/government/news/new-lung-cancer-screening-roll-out-to-detect-cancer-sooner
15. O'Dowd EL, Lee RW, Akram AR, et al. Defining the road map to a UK national lung cancer screening programme. Lancet Oncol 2023;24:e207-18.
16. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol 2015;44:827-36.
17. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-8.
18. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015;22:553-64.
19. Gilbert J, Rao G, Schuemie M, et al. Cohort diagnostics: diagnostics for OHDSI studies. 2020. Accessed 3 Jun 2024. Available online: https://ohdsi.github.io/CohortDiagnostics
20. Gallagher AM, Dedman D, Padmanabhan S, et al. The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations. Pharmacoepidemiol Drug Saf 2019;28:563-9.
21. Eurostat Taskforce. Revision of the European Standard Population - Report of Eurostat's task force. 2013. Available online: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-ra-13-028 (accessed 2nd August 2023).
22. Raventós B, Català M, Du M, et al. IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model. Pharmacoepidemiol Drug Saf 2024;33:e5717.
23. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Springer, New York. ISBN 0-387-98784-3; 2000.
24. NHS England [Internet]. 2023 [cited 2024 Jan 9]. Cancer Prevalence. National Disease Registration Service. Available online: https://www.cancerdata.nhs.uk/incidence_and_mortality
25. Cancer Research UK. Lung Cancer Statistics. Accessed 26 Feb 2024. Available online: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-

cancer-type/lung-cancer
26. Gysling S, Morgan H, Ifesemen OS, et al. The Impact of COVID-19 on Lung Cancer Incidence in England: Analysis of the National Lung Cancer Audit 2019 and 2020 Rapid Cancer Registration Datasets. Chest 2023;163:1599-607.
27. Iyen-Omofoman B, Hubbard RB, Smith CJ, et al. The distribution of lung cancer across sectors of society in the United Kingdom: a study using national primary care data. BMC Public Health 2011;11:857.
28. Ganti AK, Klein AB, Cotarla I, et al. Update of Incidence, Prevalence, Survival, and Initial Treatment in Patients With Non-Small Cell Lung Cancer in the US. JAMA Oncol 2021;7:1824-32.
29. Gridelli C, Rossi A, Carbone DP, et al. Non-small-cell lung cancer. Nat Rev Dis Primers 2015;1:15009.
30. Guarga L, Ameijide A, Marcos-Gragera R, et al. Trends in lung cancer incidence by age, sex and histology from 2012 to 2025 in Catalonia (Spain). Sci Rep 2021;11:23274.
31. Fu Y, Liu J, Chen Y, et al. Gender disparities in lung cancer incidence in the United States during 2001-2019. Sci Rep 2023;13:12581.
32. DeSantis CE, Miller KD, Dale W, et al. Cancer statistics for adults aged 85 years and older, 2019. CA Cancer J Clin 2019;69:452-67.
33. Royal College of Surgeons of England. National Lung Cancer Audit: State of the Nation Report 2023. Accessed 26 Feb 2024. Available online: https://www.lungcanceraudit.org.uk/reports-publications/nlca-state-of-the-nation-report-2023-version-3/
34. NHS England. National Disease Registration Service. Cancer Survival in England: adult, stage at diagnosis, childhood and geographical patterns. Available online: https://www.cancerdata.nhs.uk/survival/cancersurvivalengland
35. Lu T, Yang X, Huang Y, et al. Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. Cancer Manag Res 2019;11:943-53.
36. Allemani C, Matsuda T, Di Carlo V, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet 2018;391:1023-75.
37. Francisci S, Minicozzi P, Pierannunzio D, et al. Survival patterns in lung and pleural cancer in Europe 1999-2007: Results from the EUROCARE-5 study. Eur J Cancer 2015;51:2242-53.
38. Smith DR, Behzadnia A, Imawana RA, et al. Exposure-lag response of smoking prevalence on lung cancer incidence using a distributed lag non-linear model. Sci Rep 2021;11:14478.
39. Feliu A, Filippidis FT, Joossens L, et al. Impact of tobacco control policies on smoking prevalence and quit ratios in 27 European Union countries from 2006 to 2014. Tob Control 2019;28:101-9.
40. Jackson SE, Beard E, Angus C, et al. Moderators of changes in smoking, drinking and quitting behaviour associated with the first COVID-19 lockdown in England. Addiction 2022;117:772-83.
41. Tattan-Birch H, Jarvis MJ. Children's exposure to second-hand smoke 10 years on from smoke-free legislation in England: Cotinine data from the Health Survey for England 1998-2018. Lancet Reg Health Eur 2022;15:100315.
42. Bracken-Clarke D, Kapoor D, Baird AM, et al. Vaping and lung cancer - A review of current data and recommendations. Lung Cancer 2021;153:11-20.
43. Stone PW, Osen M, Ellis A, et al. Prevalence of Chronic Obstructive Pulmonary Disease in England from 2000 to 2019. Int J Chron Obstruct Pulmon Dis 2023;18:1565-74.
44. Public Health England. UK National Radon Action Plan. 2018 Dec. Accessed 26 Feb 2024. Available online: https://www.gov.uk/government/publications/uk-national-radon-action-plan
45. Aßenmacher M, Kaiser JC, Zaballa I, et al. Exposure-lag-response associations between lung cancer mortality and radon exposure in German uranium miners. Radiat Environ Biophys 2019;58:321-36.
46. Turner MC, Andersen ZJ, Baccarelli A, et al. Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. CA Cancer J Clin 2020. [Epub ahead of print]. doi: 10.3322/caac.21632.
47. Huang Y, Zhu M, Ji M, et al. Air Pollution, Genetic Factors, and the Risk of Lung Cancer: A Prospective Study in the UK Biobank. Am J Respir Crit Care Med 2021;204:817-25. Erratum in: Am J Respir Crit Care Med 2022;205:1254.
48. May L, Shows K, Nana-Sinkam P, et al. Sex Differences in Lung Cancer. Cancers (Basel) 2023;15:3111.
49. Ragavan M, Patel MI. The evolving landscape of sex-based differences in lung cancer: a distinct disease in women. Eur Respir Rev 2022;31:210100.
50. Chien LH, Jiang HF, Tsai FY, et al. Incidence of Lung Adenocarcinoma by Age, Sex, and Smoking Status in Taiwan. JAMA Netw Open 2023;6:e2340704.

51. Xiong WM, Xu QP, Li X, et al. The association between human papillomavirus infection and lung cancer: a system review and meta-analysis. Oncotarget 2017;8:96419-32.
52. Barrera-Rodriguez R, Morales-Fuentes J. Lung cancer in women. Lung Cancer (Auckl) 2012;3:79-89.
53. Mollerup S, Berge G, Baera R, et al. Sex differences in risk of lung cancer: Expression of genes in the PAH bioactivation pathway in relation to smoking and bulky DNA adducts. Int J Cancer 2006;119:741-4.
54. Stapelfeld C, Dammann C, Maser E. Sex-specificity in lung cancer risk. Int J Cancer 2020;146:2376-82.
55. Denton E, Conron M. Improving outcomes in lung cancer: the value of the multidisciplinary health care team. J Multidiscip Healthc 2016;9:137-44.
56. Pillay B, Wootten AC, Crowe H, et al. The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: A systematic review of the literature. Cancer Treat Rev 2016;42:56-72.
57. Navani N, Baldwin DR, Edwards JG, et al. Lung Cancer in the United Kingdom. J Thorac Oncol 2022;17:186-93.
58. Cai H, Wang Y, Qin D, et al. Advanced surgical technologies for lung cancer treatment: Current status and perspectives. Engineered Regeneration 2023;4:55-67.
59. Ironmonger L, Ohuma E, Ormiston-Smith N, et al. An evaluation of the impact of large-scale interventions to raise public awareness of a lung cancer symptom. Br J Cancer 2015;112:207-16.
60. Ball S, Hyde C, Hamilton W, et al. An evaluation of a national mass media campaign to raise public awareness of possible lung cancer symptoms in England in 2016 and 2017. Br J Cancer 2022;126:187-95.
61. Yang SC, Wang JD, Wang SY. Considering lead-time bias in evaluating the effectiveness of lung cancer screening with real-world data. Sci Rep 2021;11:12180.
62. Roland M. Linking physicians' pay to the quality of care--a major experiment in the United kingdom. N Engl J Med 2004;351:1448-54.
63. Swerdlow AJ. Cancer Registration in England and Wales: Some Aspects Relevant to Interpretation of the Data. J R Stat Soc 1986;149:146-60.
64. Sarfati D, Blakely T, Pearce N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. Int J Epidemiol 2010;39:598-610.
65. Arhi CS, Bottle A, Burns EM, et al. Comparison of cancer diagnosis recording between the Clinical Practice Research Datalink, Cancer Registry and Hospital Episodes Statistics. Cancer Epidemiol 2018;57:148-57.
66. Strongman H, Williams R, Bhaskaran K. 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 2020;10:e037719.
67. Mincuzzi A, Carone S, Galluzzo C, et al. Gender differences, environmental pressures, tumor characteristics, and death rate in a lung cancer cohort: a seven-years Bayesian survival analysis using cancer registry data from a contaminated area in Italy. Front Public Health 2023;11:1278416.
68. Rudin CM, Brambilla E, Faivre-Finn C, et al. Small-cell lung cancer. Nat Rev Dis Primers 2021;7:3.

Cite this article as: Corby G, Barclay NL, Tan EH, Burn E, Delmestri A, Duarte-Salles T, Golozar A, Man WY, Tietzova I, Prieto-Alhambra D, Newby D. Incidence, prevalence, and survival of lung cancer in the United Kingdom from 2000–2021: a population-based cohort study. Transl Lung Cancer Res 2024;13(9):2187-2201. doi: 10.21037/tlcr-24-241

Table S1 Clinical code lists for lung cancer

Concept ID	Concept SNOMED code	Concept description	Code used for
443388	363358000	Malignant tumor of lung	Incidence and prevalence
258369	93880001	Primary malignant neoplasm of lung	Incidence and prevalence
4092216	187862007	Malignant neoplasm of upper lobe of lung	Incidence and prevalence
4089756	187870002	Malignant neoplasm of lower lobe of lung	Incidence and prevalence
4151250	269464000	Malignant neoplasm of upper lobe, bronchus or lung	Incidence and prevalence
4246121	93827000	Primary malignant neoplasm of hilus of lung	Incidence and prevalence
4089754	187866005	Malignant neoplasm of middle lobe of lung	Incidence and prevalence
4092217	187864008	Malignant neoplasm of middle lobe, bronchus or lung	Incidence and prevalence
258375	109371002	Overlapping malignant neoplasm of bronchus and lung	Incidence and prevalence
4092218	187865009	Malignant neoplasm of middle lobe bronchus	Incidence and prevalence
256646	372112000	Primary malignant neoplasm of middle lobe, bronchus or lung	Incidence and prevalence
261236	372135002	Primary malignant neoplasm of upper lobe, bronchus or lung	Incidence and prevalence
433973	93729006	Primary malignant neoplasm of bronchus of left lower lobe	Incidence and prevalence
762426	354701000119107	Primary malignant neoplasm of left lung	Incidence and prevalence
762427	354741000119109	Primary malignant neoplasm of right lung	Incidence and prevalence
4110587	254625005	Malignant tumor of lung parenchyma	Incidence and prevalence
4110589	254629004	Large cell carcinoma of lung	Incidence and prevalence
4110590	254631008	Giant cell carcinoma of lung	Incidence and prevalence
4110591	254632001	Small cell carcinoma of lung	Incidence and prevalence
4110705	254634000	Squamous cell carcinoma of lung	Incidence and prevalence
4110706	254638002	Pancoast tumor	Incidence and prevalence
4111807	254635004	Epithelioid hemangioendothelioma of lung	Incidence and prevalence
4112738	254626006	Adenocarcinoma of lung	Incidence and prevalence
4112739	254633006	Oat cell carcinoma of lung	Incidence and prevalence
4115276	254637007	Non-small cell lung cancer	Incidence and prevalence
4140471	427038005	Epidermal growth factor receptor negative non-small cell lung cancer	Incidence and prevalence
4143825	426964009	Epidermal growth factor receptor positive non-small cell lung cancer	Incidence and prevalence
4155293	372111007	Carcinoma of lower lobe, bronchus or lung	Incidence and prevalence
4157454	372110008	Primary malignant neoplasm of lower lobe, bronchus or lung	Incidence and prevalence
4162248	372113005	Carcinoma of middle lobe, bronchus or lung	Incidence and prevalence
4162252	372136001	Carcinoma of upper lobe, bronchus or lung	Incidence and prevalence
4196724	313353007	Squamous cell carcinoma of bronchus in left lower lobe	Incidence and prevalence
4196725	313355000	Squamous cell carcinoma of bronchus in right lower lobe	Incidence and prevalence
4197581	313354001	Squamous cell carcinoma of bronchus in left upper lobe	Incidence and prevalence
4197582	313356004	Squamous cell carcinoma of bronchus in right middle lobe	Incidence and prevalence
4197583	313357008	Squamous cell carcinoma of bronchus in right upper lobe	Incidence and prevalence
4208307	440173001	Nonsquamous non-small cell neoplasm of lung	Incidence and prevalence
4246027	93730001	Primary malignant neoplasm of bronchus of left upper lobe	Incidence and prevalence
4246126	93865007	Primary malignant neoplasm of left upper lobe of lung	Incidence and prevalence
4246148	93991009	Primary malignant neoplasm of right lower lobe of lung	Incidence and prevalence
4246804	93732009	Primary malignant neoplasm of bronchus of right middle lobe	Incidence and prevalence
4246805	93733004	Primary malignant neoplasm of bronchus of right upper lobe	Incidence and prevalence
4247727	93731002	Primary malignant neoplasm of bronchus of right lower lobe	Incidence and prevalence
4247832	93864006	Primary malignant neoplasm of lower lobe of left lung	Incidence and prevalence
4307118	423050000	Large cell carcinoma of lung, TNM stage 2	Incidence and prevalence
4308479	423121009	Non-small cell carcinoma of lung, TNM stage 4	Incidence and prevalence
4310448	423295000	Squamous cell carcinoma of lung, TNM stage 1	Incidence and prevalence
4310703	424132000	Non-small cell carcinoma of lung, TNM stage 1	Incidence and prevalence
4311501	93992002	Primary malignant neoplasm of right middle lobe of lung	Incidence and prevalence
4311997	422968005	Non-small cell carcinoma of lung, TNM stage 3	Incidence and prevalence
4312274	425376008	Squamous cell carcinoma of lung, TNM stage 4	Incidence and prevalence
4312567	93993007	Primary malignant neoplasm of upper lobe of right lung	Incidence and prevalence
4312768	424970000	Large cell carcinoma of lung, TNM stage 3	Incidence and prevalence
4313200	423468007	Squamous cell carcinoma of lung, TNM stage 2	Incidence and prevalence
4313751	423600008	Large cell carcinoma of lung, TNM stage 4	Incidence and prevalence
4314040	424938000	Large cell carcinoma of lung, TNM stage 1	Incidence and prevalence
4314172	425048006	Non-small cell carcinoma of lung, TNM stage 2	Incidence and prevalence
4322387	425230006	Squamous cell carcinoma of lung, TNM stage 3	Incidence and prevalence
36686537	12235561000119100	Large cell carcinoma of left lung	Incidence and prevalence
36686538	12235601000119100	Large cell carcinoma of right lung	Incidence and prevalence
36712707	1078881000119100	Primary adenocarcinoma of lower lobe of left lung	Incidence and prevalence
36712708	1078901000119100	Primary adenocarcinoma of upper lobe of left lung	Incidence and prevalence
36712709	1078931000119100	Primary adenocarcinoma of lower lobe of right lung	Incidence and prevalence
36712815	12240951000119100	Squamous cell carcinoma of left lung	Incidence and prevalence
36712816	12240991000119100	Squamous cell carcinoma of right lung	Incidence and prevalence
36712981	15956381000119100	Adenocarcinoma of right lung	Incidence and prevalence
36713366	683991000119103	Extensive stage primary small cell carcinoma of lung	Incidence and prevalence
36716426	722425009	Reactive oxygen species 1 positive non-small cell lung cancer	Incidence and prevalence
36716500	722528008	Primary malignant neuroendocrine neoplasm of lung	Incidence and prevalence
36717017	1078961000119100	Primary adenocarcinoma of upper lobe of right lung	Incidence and prevalence
37109576	723301009	Squamous non-small cell lung cancer	Incidence and prevalence
37110031	724056005	Malignant neoplasm of lower lobe of right lung	Incidence and prevalence
37110032	724058006	Malignant neoplasm of upper lobe of left lung	Incidence and prevalence
37110033	724059003	Malignant neoplasm of lower lobe of left lung	Incidence and prevalence
37110034	724060008	Malignant neoplasm of right upper lobe of lung	Incidence and prevalence
37311684	822969007	Acinar cell cystadenocarcinoma of lung	Incidence and prevalence
37395648	67811000119102	Primary small cell malignant neoplasm of lung, TNM stage 1	Incidence and prevalence
37395649	67821000119109	Primary small cell malignant neoplasm of lung, TNM stage 2	Incidence and prevalence
37395650	67831000119107	Primary small cell malignant neoplasm of lung, TNM stage 3	Incidence and prevalence
37395651	67841000119103	Primary small cell malignant neoplasm of lung, TNM stage 4	Incidence and prevalence
40391740	189815007	Pulmonary blastoma	Incidence and prevalence
40492938	448993007	Carcinoma of lung	Incidence and prevalence
42539251	15956341000119100	Adenocarcinoma of left lung	Incidence and prevalence
45766129	703228009	Non-small cell lung cancer with mutation in epidermal growth factor receptor	Incidence and prevalence
45766131	703230006	Non-small cell lung cancer without mutation in epidermal growth factor receptor	Incidence and prevalence
45768879	707403002	Primary fetal adenocarcinoma of lung	Incidence and prevalence
45768880	707404008	Primary mixed subtype adenocarcinoma of lung	Incidence and prevalence
45768881	707405009	Primary adenosquamous carcinoma of lung	Incidence and prevalence
45768883	707408006	Primary small cell non-keratinizing squamous cell carcinoma of lung	Incidence and prevalence
45768884	707409003	Primary acinar cell carcinoma of lung	Incidence and prevalence
45768885	707410008	Primary solid carcinoma of lung	Incidence and prevalence
45768886	707411007	Primary papillary adenocarcinoma of lung	Incidence and prevalence
45768916	707451005	Primary adenocarcinoma of lung	Incidence and prevalence
45768917	707452003	Primary mucinous adenocarcinoma of lung	Incidence and prevalence
45768918	707453008	Primary clear cell squamous cell carcinoma of lung	Incidence and prevalence
45768919	707454002	Primary basaloid squamous cell carcinoma of lung	Incidence and prevalence
45768920	707456000	Primary undifferentiated carcinoma of lung	Incidence and prevalence
45768921	707457009	Primary spindle cell carcinoma of lung	Incidence and prevalence
45768922	707458004	Primary pleomorphic carcinoma of lung	Incidence and prevalence
45768923	707460002	Primary pseudosarcomatous carcinoma of lung	Incidence and prevalence
45768927	707464006	Primary myoepithelial carcinoma of lung	Incidence and prevalence
45768928	707466008	Primary adenoid cystic carcinoma of lung	Incidence and prevalence
45768929	707467004	Primary salivary gland type carcinoma of lung	Incidence and prevalence
45768930	707468009	Primary mixed mucinous and non-mucinous bronchiolo-alveolar carcinoma of lung	Incidence and prevalence
45768931	707469001	Primary non-mucinous bronchiolo-alveolar carcinoma of lung	Incidence and prevalence
45768932	707470000	Primary mucinous bronchiolo-alveolar carcinoma of lung	Incidence and prevalence
45769034	707595001	Primary mucinous cystadenocarcinoma of lung	Incidence and prevalence
45769035	707596000	Primary carcinosarcoma of lung	Incidence and prevalence
45772933	707407001	Primary signet ring cell carcinoma of lung	Incidence and prevalence
45772938	707455001	Primary papillary squamous cell carcinoma of lung	Incidence and prevalence
45772939	707465007	Primary mucoepidermoid carcinoma of lung	Incidence and prevalence
46272955	711414003	Primary clear cell adenocarcinoma of lung	Incidence and prevalence
4198434	314954002	Local recurrence of malignant tumor of lung	Prevalence only
4201621	315006004	Metastasis from malignant tumor of lung	Prevalence only

The clinical codelists used for lung cancer is listed in the table below with the corresponding SNOMED concept ID, OMOP concept ID and concept description. Only diagnosis records alone were used to identify cancer outcome for this study. Different codelists were created for incident and prevalent definitions of lung cancer. We developed concept definitions using ATLAS, the OHDSI open-source platform (https://github.com/OHDSI/atlas). Clinical adjudicators reviewed the cohort definitions and associated concept sets.

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

Table S2 Population attrition showing eligible patients for study from each database

N	Reason	Excluded, n	Database
39999011	Starting population		Aurum
39999011	Missing year of birth	0	
39999011	Missing sex	0	
34833388	Cannot satisfy age criteria during the study period based on year of birth	5165623	
29190480	No observation time available during study period	5642908	
29190480	Doesn't satisfy age criteria during the study period	0	
25483313	Prior history requirement not fulfilled during study period	3707167	
24340860	No observation time available after applying age and prior history criteria	1142453	
24340860	Starting analysis population		
24340860	Estimating prevalence		
24334950	Excluded due to prior event (do not pass outcome washout during study period)	5910	
24334950	Estimating incidence		
88540	With a cancer diagnosis	24246410	
86710	Cancer diagnosis not on same date as death	1830	
86710	Estimating survival		
17054819	Starting population		GOLD
17054819	Missing year of birth	0	
17054819	Missing sex	0	
15210165	Cannot satisfy age criteria during the study period based on year of birth	1844654	
13978229	No observation time available during study period	1231936	
13978229	Doesn't satisfy age criteria during the study period	0	
12254874	Prior history requirement not fulfilled during study period	1723355	
11388117	No observation time available after applying age and prior history criteria	866757	
11388117	Starting analysis population		
11388117	Estimating prevalence		
11385621	Excluded due to prior event (do not pass outcome washout during study period)	2496	
11385621	Estimating incidence		
45563	With a cancer diagnosis	11340058	
43903	Cancer diagnosis not on same date as death	1660	
43903	Estimating survival		

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

Table S3 Baseline characteristics of lung cancer patients at the time of diagnosis for CPRD GOLD stratified by region (England, Northern Ireland, Scotland and Wales)

Region	England	Northern Ireland	Scotland	Wales
Number of lung cancer patients	21595	2363	14118	7487
Sex: male	12087 (56.0%)	1273 (53.9%)	7202 (51.0%)	4007 (53.5%)
Age (years), median (IQR)	73 (65 to 79)	71 (64 to 78)	72 (65 to 79)	73 (65 to 79)
Age groups (years), N (%)				
18-29	16 (0.1%)	0	<5	5 (0.1%)
30-39	63 (0.3%)	<5	49 (0.3%)	15 (0.2%)
40-49	519 (2.4%)	55 (2.3%)	299 (2.1%)	157 (2.1%)
50-59	2212 (10.2%)	260 (11.0%)	1484 (10.5%)	704 (9.4%)
60-69	5692 (26.4%)	710 (30.0%)	3857 (27.3%)	1990 (26.6%)
70-79	7750 (35.9%)	843 (35.7%)	5317 (37.7%)	2835 (37.9%)
80-89	4706 (21.8%)	447 (18.9%)	2820 (20.0%)	1573 (21.0%)
90+	637 (2.9%)	46 (1.9%)	289 (2.0%)	208 (2.8%)
Prior history, days	3691 (2,057 to 5,312)	4180 (2,419 to 5,788)	3444.5 (1,854 to 5,193)	3820 (2,052 to 5,642)
Smoking status (any time 5 years prior)				
Non-smoker	3600 (16.7%)	410 (17.4%)	1896 (13.4%)	1248 (16.7%)
Former smoker	367 (1.7%)	19 (0.8%)	92 (0.7%)	65 (0.9%)
Current smoker	9331 (43.2%)	1321 (55.9%)	7673 (54.3%)	3694 (49.3%)
Missing	8297 (38.4%)	613 (25.9%)	4457 (31.6%)	2480 (33.1%)
General conditions (any time prior)				
Atrial Fibrillation	1535 (7.1%)	192 (8.1%)	904 (6.4%)	559 (7.5%)
Cerebrovascular Disease	1806 (8.4%)	212 (9.0%)	1207 (8.5%)	600 (8.0%)
Chronic Liver Disease	86 (0.4%)	19 (0.8%)	89 (0.6%)	48 (0.6%)
Chronic Obstructive Lung Disease	5058 (23.4%)	650 (27.5%)	3537 (25.1%)	1879 (25.1%)
Coronary Arteriosclerosis	292 (1.4%)	20 (0.8%)	274 (1.9%)	89 (1.2%)
Crohn's Disease	65 (0.3%)	16 (0.7%)	48 (0.3%)	25 (0.3%)
Dementia	332 (1.5%)	52 (2.2%)	270 (1.9%)	111 (1.5%)
Depressive Disorder	3237 (15.0%)	455 (19.3%)	1716 (12.2%)	981 (13.1%)
Diabetes Mellitus	2474 (11.5%)	311 (13.2%)	1553 (11.0%)	967 (12.9%)
Gastroesophageal Reflux Disease	513 (2.4%)	56 (2.4%)	404 (2.9%)	283 (3.8%)
Gastrointestinal Hemorrhage	1691 (7.8%)	235 (9.9%)	614 (4.3%)	559 (7.5%)
Heart Disease	5305 (24.6%)	598 (25.3%)	3079 (21.8%)	1698 (22.7%)
Heart Failure	1075 (5.0%)	114 (4.8%)	521 (3.7%)	291 (3.9%)
HIV	5 (0.0%)	<5	<5	<5
Hyperlipidemia	2367 (11.0%)	350 (14.8%)	1011 (7.2%)	853 (11.4%)
Hypertensive disorder	6293 (29.1%)	639 (27.0%)	3363 (23.8%)	2090 (27.9%)
Ischemic Heart Disease	3180 (14.7%)	356 (15.1%)	1746 (12.4%)	946 (12.6%)
Obesity	432 (2.0%)	64 (2.7%)	251 (1.8%)	181 (2.4%)
Osteoarthritis	5218 (24.2%)	491 (20.8%)	2499 (17.7%)	1622 (21.7%)
Peripheral Vascular Disease	1014 (4.7%)	151 (6.4%)	789 (5.6%)	301 (4.0%)
Pneumonia	1232 (5.7%)	168 (7.1%)	727 (5.1%)	406 (5.4%)
Pulmonary Embolism	386 (1.8%)	29 (1.2%)	242 (1.7%)	135 (1.8%)
Renal Impairment	2832 (13.1%)	389 (16.5%)	1823 (12.9%)	1086 (14.5%)
Ulcerative Colitis	103 (0.5%)	13 (0.6%)	36 (0.3%)	29 (0.4%)
Venous Thrombosis	1275 (5.9%)	134 (5.7%)	482 (3.4%)	427 (5.7%)

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

Table S4 Baseline characteristics of lung cancer patients at the time of diagnosis for CPRD GOLD stratified by smoking status (any time 5 years prior)

Smoking status	Non-smoker	Former smoker	Smoker	Missing
Number of lung cancer patients	7154	543	22019	15847
Sex: male	3487 (48.7%)	315 (58.0%)	11446 (52.0%)	9321 (58.8%)
Age (years), median (IQR)	77 (69 to 83)	72 (65 to 77)	69 (62 to 76)	75 (68 to 81)
Age groups (years), N (%)				
18-29	14 (0.2%)	0	<5	6 (0.0%)
30-39	38 (0.5%)	<5	61 (0.3%)	29 (0.2%)
40-49	111 (1.6%)	9 (1.7%)	681 (3.1%)	229 (1.4%)
50-59	377 (5.3%)	37 (6.8%)	3109 (14.1%)	1137 (7.2%)
60-69	1263 (17.7%)	182 (33.5%)	7436 (33.8%)	3368 (21.3%)
70-79	2536 (35.4%)	231 (42.5%)	7678 (34.9%)	6300 (39.8%)
80-89	2366 (33.1%)	75 (13.8%)	2885 (13.1%)	4220 (26.6%)
90+	449 (6.3%)	8 (1.5%)	165 (0.7%)	558 (3.5%)
Prior history, days	3626.5 (1,978 to 5,432)	3198 (1,691 to 4,836)	3676 (2,105 to 5,357)	3662 (1,883 to 5,320)
General conditions (any time prior)				
Atrial Fibrillation	687 (9.6%)	36 (6.6%)	1119 (5.1%)	1348 (8.5%)
Cerebrovascular Disease	606 (8.5%)	47 (8.7%)	1812 (8.2%)	1360 (8.6%)
Chronic Liver Disease	25 (0.3%)	7 (1.3%)	153 (0.7%)	57 (0.4%)
Chronic Obstructive Lung Disease	905 (12.7%)	181 (33.3%)	6588 (29.9%)	3450 (21.8%)
Coronary Arteriosclerosis	114 (1.6%)	7 (1.3%)	304 (1.4%)	250 (1.6%)
Crohn's Disease	19 (0.3%)	<5	87 (0.4%)	47 (0.3%)
Dementia	155 (2.2%)	<5	305 (1.4%)	302 (1.9%)
Depressive Disorder	793 (11.1%)	64 (11.8%)	3721 (16.9%)	1811 (11.4%)
Diabetes Mellitus	1101 (15.4%)	73 (13.4%)	2406 (10.9%)	1725 (10.9%)
Gastroesophageal Reflux Disease	229 (3.2%)	23 (4.2%)	556 (2.5%)	448 (2.8%)
Gastrointestinal Hemorrhage	590 (8.2%)	39 (7.2%)	1353 (6.1%)	1117 (7.0%)
Heart Disease	1979 (27.7%)	148 (27.3%)	4422 (20.1%)	4131 (26.1%)
Heart Failure	368 (5.1%)	31 (5.7%)	707 (3.2%)	895 (5.6%)
HIV	<5	<5	7 (0.0%)	<5
Hyperlipidemia	823 (11.5%)	62 (11.4%)	2227 (10.1%)	1469 (9.3%)
Hypertensive disorder	2420 (33.8%)	152 (28.0%)	5587 (25.4%)	4226 (26.7%)
Ischemic Heart Disease	1115 (15.6%)	103 (19.0%)	2728 (12.4%)	2282 (14.4%)
Obesity	186 (2.6%)	11 (2.0%)	391 (1.8%)	340 (2.1%)
Osteoarthritis	1829 (25.6%)	117 (21.5%)	4163 (18.9%)	3721 (23.5%)
Peripheral Vascular Disease	208 (2.9%)	30 (5.5%)	1332 (6.0%)	685 (4.3%)
Pneumonia	377 (5.3%)	37 (6.8%)	1156 (5.3%)	963 (6.1%)
Pulmonary Embolism	170 (2.4%)	<5	309 (1.4%)	309 (1.9%)
Renal Impairment	1466 (20.5%)	67 (12.3%)	2353 (10.7%)	2244 (14.2%)
Ulcerative Colitis	34 (0.5%)	<5	60 (0.3%)	83 (0.5%)
Venous Thrombosis	463 (6.5%)	32 (5.9%)	908 (4.1%)	915 (5.8%)

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

Table S5 Baseline characteristics of lung cancer patients at the time of diagnosis for CPRD Aurum

Database	CPRD Aurum
Number of patients	88540
Sex: male	48819 (55.1%)
Age, median (IQR)	73 (65 to 80)
Age groups, years	
18-29	45 (0.1%)
30-39	274 (0.3%)
40-49	1880 (2.1%)
50-59	8981 (10.1%)
60-69	23278 (26.3%)
70-79	31897 (36.0%)
80-89	19488 (22.0%)
90+	2697 (3.00%)
Prior history, days, median [IQR]	6775 (3,487 to 10,987)
General conditions (any time prior)	
Atrial fibrillation	7056 (8.0%)
Cerebrovascular disease	8298 (9.4%)
Chronic liver disease	530 (0.6%)
Chronic obstructive lung disease	23599 (26.7%)
Coronary arteriosclerosis	1261 (1.4%)
Crohn's disease	346 (0.4%)
Dementia	1656 (1.9%)
Depressive disorder	12974 (14.7%)
Diabetes mellitus	12822 (14.5%)
Gastroesophageal reflux disease	3058 (3.5%)
Gastrointestinal haemorrhage	6691 (7.6%)
Heart disease	24166 (27.3%)
Heart failure	4276 (4.8%)
Hepatitis C	126 (0.1%)
HIV	23 (0.00%)
Hyperlipidemia	10449 (11.8%)
Hypertensive disorder	35469 (40.1%)
Ischemic heart disease	14929 (16.9%)
Lesion of liver	1556 (1.8%)
Obesity	2403 (2.7%)
Osteoarthritis	25083 (28.3%)
Peripheral vascular disease	4932 (5.6%)
Pneumonia	6741 (7.6%)
Psoriasis	3740 (4.2%)
Pulmonary embolism	1923 (2.2%)
Renal impairment	13373 (15.1%)
Rheumatoid arthritis	2104 (2.4%)
Schizophrenia	420 (0.5%)
Ulcerative colitis	472 (0.5%)
Urinary tract infectious disease	12018 (13.6%)
Venous thrombosis	5252 (5.9%)
Visual system disorder	36552 (41.3%)

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

Table S6 Overall incidence rates (with 95% confidence intervals) for lung cancer from 2000 to 2021 for CPRD GOLD and 2000 to 2019 for Aurum stratified by database and age group

Age group (years)	People, n	pys	Events, n	Incidence (100,000 pys)	Database
18 to 29	9,238,441	31,941,436	45	0.14 (0.10 to 0.19)	CPRD Aurum
30 to 39	8,292,622	32,680,100	274	0.84 (0.74 to 0.94)	
40 to 49	6,515,966	32,924,342	1,880	5.71 (5.45 to 5.97)	
50 to 59	5,438,224	28,668,826	8,981	31.33 (30.7 to 32.0)	
60 to 69	4,170,276	22,500,850	23,278	103.5 (102.1 to 104.8)	
70 to 79	3,119,104	16,252,188	31,897	196.3 (194.1 to 198.4)	
80 to 89	1,949,102	8,834,297	19,488	220.59 (217.5 to 223.7)	
90+	666,317	2,269,618	2,697	118.8 (114.4 to 123.4)	
18 to 29	3,871,131	16,018,568	24	0.2 (0.1 to 0.2)	CPRD GOLD
30 to 39	3,682,403	15,107,431	129	0.9 (0.7 to 1.0)	
40 to 49	3,246,901	16,116,689	1,030	6.4 (6.0 to 6.8)	
50 to 59	2,885,599	14,721,668	4,660	31.7 (30.8 to 32.6)	
60 to 69	2,293,348	11,892,162	12,249	103.0 (101.2 to 104.8)	
70 to 79	1,682,118	8,407,711	16,745	199.2 (196.2 to 202.2)	
80 to 89	1,023,695	4,422,107	9,546	215.9 (211.6 to 220.2)	
90+	322,801	965,457	1,181	122.3 (115.5 to 129.5)	

Pys, person years.

Table S7 Survival (%) after 1, 5 and 10 years after lung cancer diagnosis stratified by database and sex

Time (years)	Sex	% Survival (95% CI)	Database
1	Male	38.0 (37.6–38.5)	Aurum
5		11.1 (10.7–11.4)	
10		5.6 (5.3–5.9)	
1	Female	44.6 (44.1–45.2)	
5		16.0 (15.6–16.5)	
10		9.46 (9.0–9.9)	
1	Male	35.8 (35.21–36.48)	GOLD
5		10.0 (9.6–10.5)	
10		5.0 (4.6–5.4)	
1	Female	42.6 (41.9–43.4)	
5		14.4 (13.8–15.0)	
10		8.2 (7.6–8.78)	

CI, confidence interval.

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

Table S8 Median survival stratified by database and age group

Age group (years)	Median survival in years (95% CI)	People, n	Events, n	Database
18 to 29	Not achieved	45	14	Aurum
30 to 39	1.314 (1.008 - 2.062)	272	163	
40 to 49	0.947 (0.862 - 1.038)	1850	1323	
50 to 59	0.917 (0.890 - 0.945)	8845	6603	
60 to 69	0.851 (0.830 - 0.871)	22923	17474	
70 to 79	0.698 (0.684 - 0.715)	31247	24582	
80 to 89	0.504 (0.485 - 0.523)	18932	15166	
90 +	0.361 (0.329 - 0.397)	2596	2096	
18 to 29	Not achieved	24	6	GOLD
30 to 39	1.478 (0.810 - 2.177)	127	82	
40 to 49	0.903 (0.786 - 1.027)	999	740	
50 to 59	0.827 (0.789 - 0.873)	4549	3467	
60 to 69	0.797 (0.775 - 0.824)	11907	9415	
70 to 79	0.638 (0.621 - 0.657)	16164	13232	
80 to 89	0.463 (0.449 - 0.479)	9041	7535	
90 +	0.353 (0.301 - 0.386)	1092	904	

Not achieved: median survival was not achieved in study period.

Table S9 Survival rates (and 95% confidence intervals) of lung cancer from for GOLD (2000-2021) and Aurum (2000-2019) stratified by database and age group

Age group (years)	One-year survival (%)		Five-year survival (%)		Ten-year survival (%)	
	GOLD	Aurum	GOLD	Aurum	GOLD	Aurum
18-29	82.7 (68.6–99.7)	77.3 (65.3–91.7)	71.6 (54.3–94.2)	57.5 (40.4–81.6)	–	–
30-39	54.0 (45.7–63.7)	56.0 (50.1–62.6)	30.3 (22.6–40.6)	31.9 (26.1–39.1)	21.6 (13.1–35.4)	26.0 (20.1–33.5)
40-49	47.8 (44.7–51.1)	48.6 (46.3–51.1)	19.4 (16.8–22.4)	21.3 (19.3–23.6)	14.8 (12.2–18.0)	15.5 (13.4–17.9)
50-59	44.7 (43.2–46.2)	47.2 (46.1–48.3)	16.4 (15.2–17.7)	17.6 (16.7–18.5)	11.3 (10.1–12.6)	12.0 (11.1–13.0)
60-69	43.5 (42.6–44.5)	45.4 (44.7–46.0)	15.1 (14.3–15.8)	16.8 (16.2–17.3)	8.6 (7.9–9.4)	10.1 (9.49–10.7)
70-79	38.4 (37.6–39.2)	40.9 (40.3–41.4)	11.3 (10.7–11.9)	12.6 (12.1–13.0)	5.1 (4.5–5.6)	5.76 (5.34–6.22)
80-89	31.5 (30.5–32.5)	33.7 (33.0–34.4)	6.2 (5.6–6.9)	7.5 (7.0–8.0)	1.5 (1.0–2.1)	2.25 (1.83–2.75)
90+	24.2 (21.6–27.0)	25.6 (23.9–27.5)	3.2 (1.9–5.5)	2.53 (1.70–3.78)	–	0.51 (0.17–1.50)

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

Table S10 Median survival stratified by database, calendar year for whole population and sex

Sex	Calendar year	Median survival in years (95% CI)	People, n	Events, n	Database
Both	2000 to 2004	0.635 (0.613–0.654)	12019	8028	Aurum
Male		0.591 (0.564–0.616)	7188	4977	
Female		0.701 (0.671–0.753)	4831	3051	
Both	2005 to 2009	0.594 (0.578–0.608)	20204	14397	
Male		0.561 (0.542–0.583)	11592	8529	
Female		0.646 (0.616–0.671)	8612	5868	
Both	2010 to 2014	0.690 (0.674–0.709)	25364	16950	
Male		0.624 (0.602–0.649)	13678	9458	
Female		0.772 (0.745–0.799)	11686	7492	
Both	2015 to 2019	0.923 (0.898–0.947)	29123	17285	
Male		0.783 (0.756–0.819)	15311	9670	
Female		1.125 (1.073–1.175)	13812	7615	
Both	2000 to 2004	0.559 (0.526–0.589)	6250	4259	GOLD
Male		0.512 (0.482–0.548)	3718	2617	
Female		0.630 (0.586–0.679)	2532	1642	
Both	2005 to 2009	0.594 (0.578–0.613)	12111	8617	
Male		0.553 (0.534–0.578)	6792	4963	
Female		0.660 (0.627–0.687)	5319	3654	
Both	2010 to 2014	0.668 (0.643–0.693)	12847	8616	
Male		0.611 (0.591–0.643)	6780	4691	
Female		0.745 (0.706–0.786)	6067	3925	
Both	2015 to 2019	0.832 (0.791–0.865)	9650	5951	
Male		0.709 (0.665–0.745)	4808	3099	
Female		0.950 (0.898–1.021)	4842	2852	
Both	2020 to 2021	0.767 (0.690–0.810)	3045	1490	
Male		0.712 (0.619–0.789)	1523	776	
Female		0.816 (0.709–0.958)	1522	714	

CI, confidence interval.

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

Table S11 Survival (%) after 1 and 5 years after lung cancer diagnosis stratified by database, sex and calendar year

Calendar year	Time (years)	% Survival (95% CI)	Sex	Database
2000 to 2004	1	37.23 (36.29–38.20)	Both	Aurum
2005 to 2009	1	35.17 (34.46–35.89)		
2010 to 2014	1	40.06 (39.40–40.72)		
2015 to 2019	1	48.02 (47.40–48.65)		
2000 to 2004	1	33.43 (32.14–34.77)		GOLD
2005 to 2009	1	35.28 (34.36–36.21)		
2010 to 2014	1	39.58 (38.67–40.51)		
2015 to 2019	1	45.20 (44.13–46.30)		
2020 to 2021	1	44.12 (42.06–46.28)		
2000 to 2004	1	40.23 (38.72–41.80)	Female	Aurum
2005 to 2009	1	37.95 (36.84–39.09)		
2010 to 2014	1	43.11 (42.14–44.10)		
2015 to 2019	1	52.41 (51.52–53.32)		
2000 to 2004	1	37.38 (35.32–39.56)		GOLD
2005 to 2009	1	38.83 (37.43–40.27)		
2010 to 2014	1	42.64 (41.32–44.01)		
2015 to 2019	1	48.98 (47.49–50.53)		
2020 to 2021	1	46.26 (43.37–49.35)		
2000 to 2004	1	35.24 (34.04–36.48)	Male	Aurum
2005 to 2009	1	33.14 (32.23–34.09)		
2010 to 2014	1	37.45 (36.57–38.34)		
2015 to 2019	1	44.07 (43.22–44.93)		
2000 to 2004	1	30.77 (29.14–32.48)		GOLD
2005 to 2009	1	32.49 (31.30–33.73)		
2010 to 2014	1	36.82 (35.59–38.09)		
2015 to 2019	1	41.29 (39.78–42.85)		
2020 to 2021	1	41.99 (39.12–45.07)		
2000 to 2004	5	10.47 (9.24–11.85)	Both	Aurum
2005 to 2009	5	9.09 (8.25–10.02)		
2010 to 2014	5	12.53 (11.56–13.58)		
2015 to 2019	5	19.51 (18.33–20.77)		
2000 to 2004	5	8.89 (7.30–10.82)		GOLD
2005 to 2009	5	8.46 (7.00–10.22)		
2010 to 2014	5	12.96 (11.84–14.19)		
2015 to 2019	5	17.31 (15.81–18.94)		
2000 to 2004	5	12.91 (10.73–15.54)	Female	Aurum
2005 to 2009	5	11.03 (9.70–12.54)		
2010 to 2014	5	15.12 (13.72–16.67)		
2015 to 2019	5	24.41 (22.84–26.09)		
2000 to 2004	5	10.39 (7.52–14.34)		GOLD
2005 to 2009	5	8.90 (6.42–12.34)		
2010 to 2014	5	15.43 (13.68–17.40)		
2015 to 2019	5	21.06 (18.80–23.58)		
2000 to 2004	5	8.94 (7.56–10.57)	Male	Aurum
2005 to 2009	5	7.69 (6.63–8.91)		
2010 to 2014	5	10.31 (9.02–11.79)		
2015 to 2019	5	15.05 (13.24–17.10)		
2000 to 2004	5	7.72 (6.06–9.85)		GOLD
2005 to 2009	5	8.24 (6.95–9.76)		
2010 to 2014	5	10.69 (9.28–12.31)		
2015 to 2019	5	13.49 (11.64–15.64)		

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

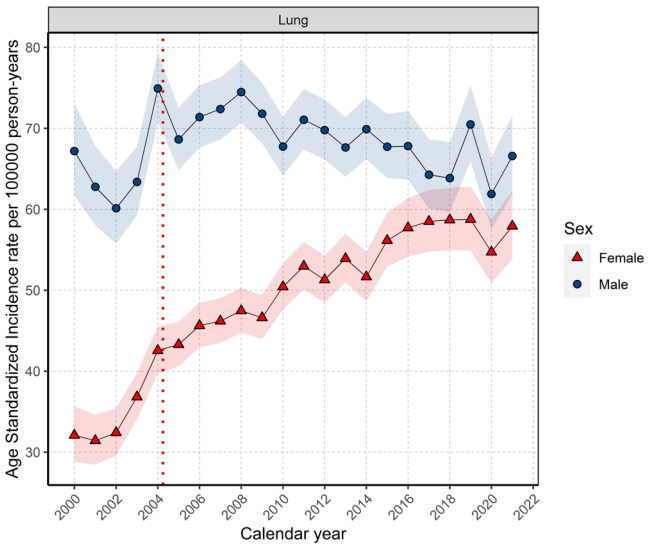

Figure S1 Annualised age standardized incidence rates for lung cancer from 2000 to 2021 for CPRD GOLD stratified by sex-age standardized to the European Standard Population 2013. Band represents 95% CI.

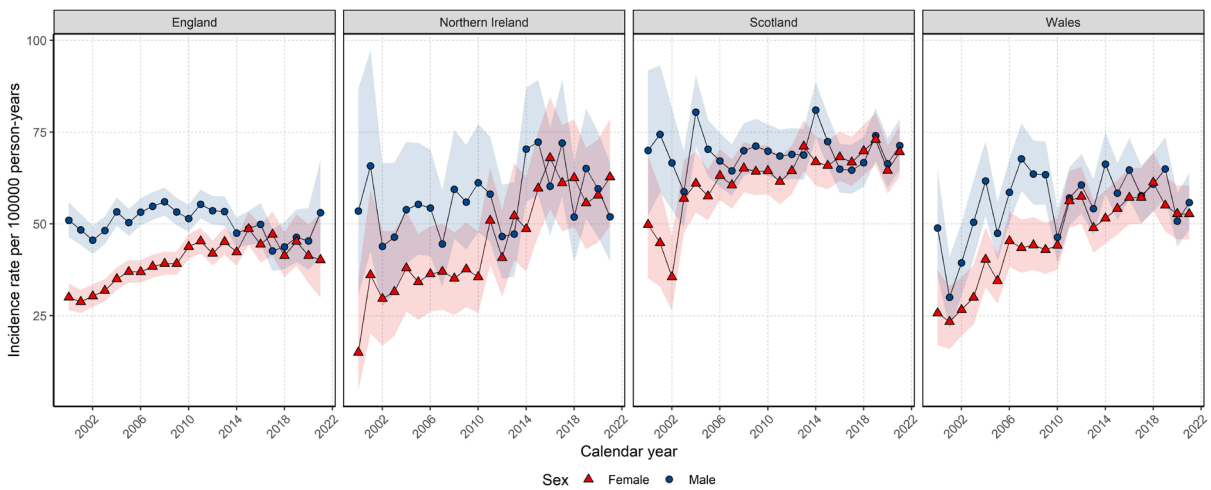

Figure S2 Annualised age standardized incidence rates for lung cancer from 2000 to 2021 for CPRD GOLD stratified by sex and region-age standardized to the European Standard Population 2013. Band represents 95% CI.

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

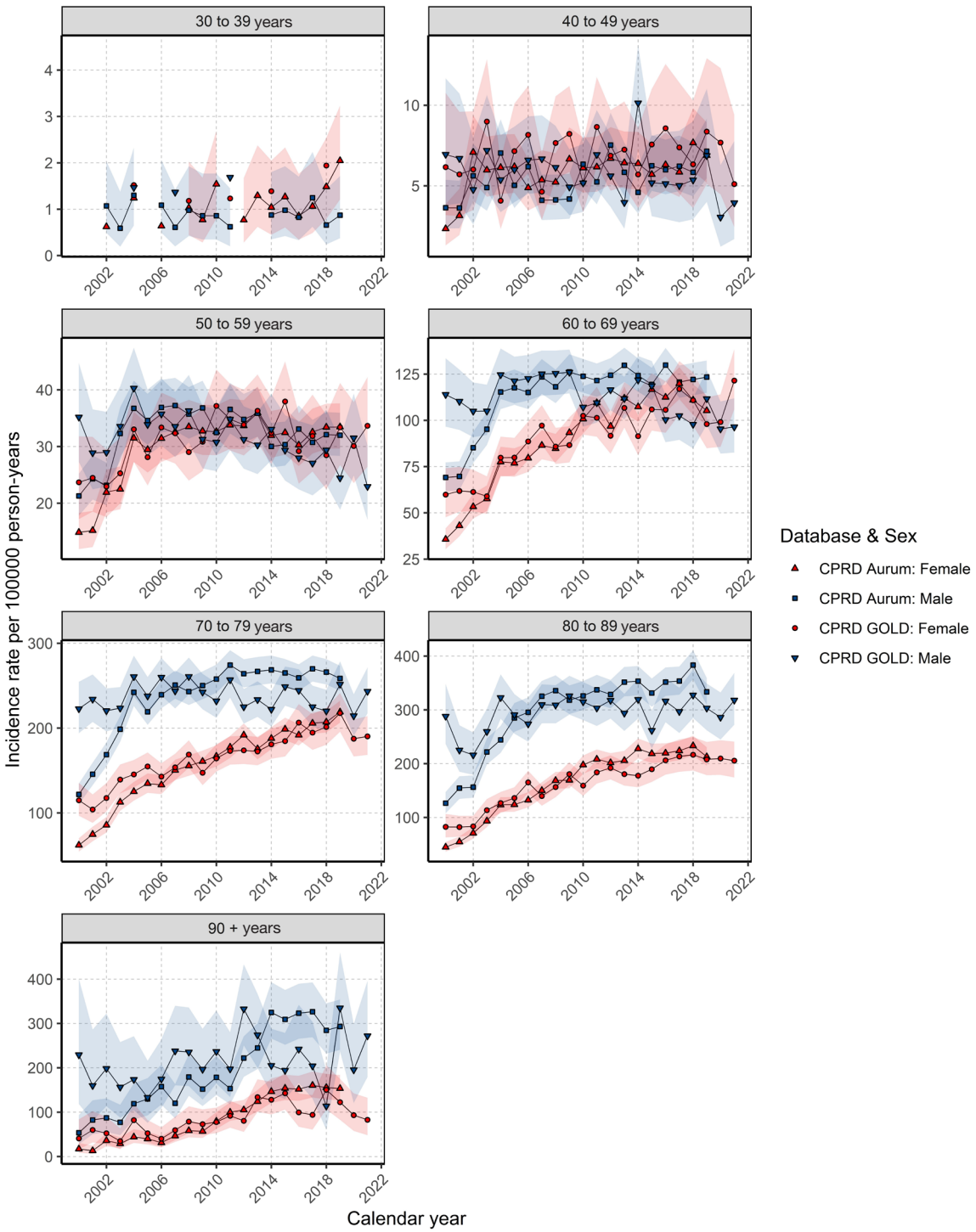

Figure S3 Annualised incidence rates for lung cancer from 2000 to 2021 stratified by database, sex and age group. Band represents 95% CI.

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

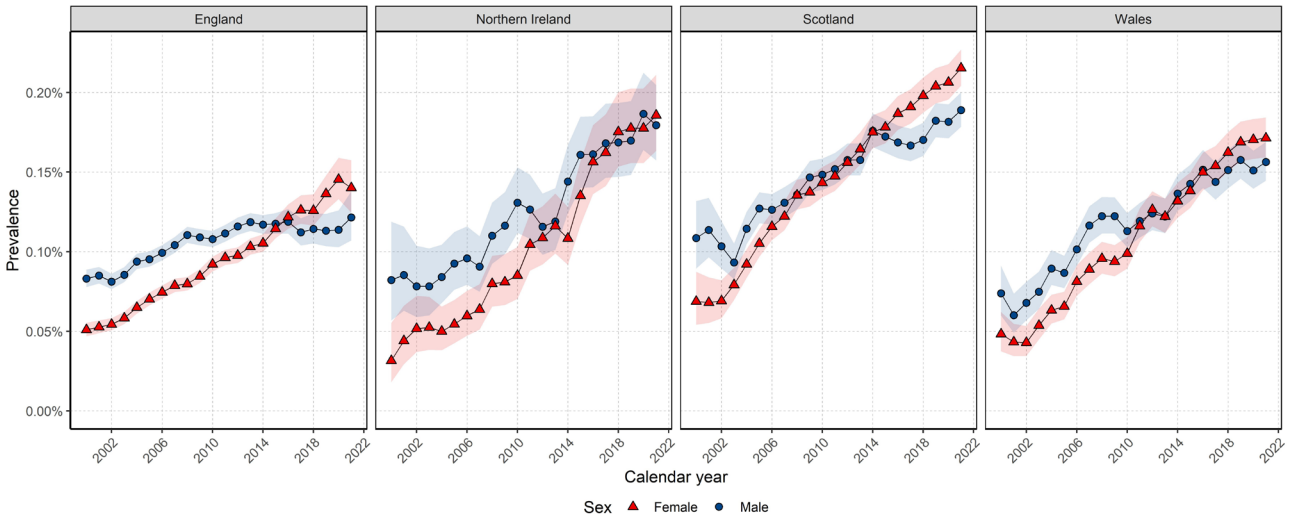

Figure S4 Annual period prevalence from 2000 to 2021 stratified by UK region. Bands show 95% confidence interval.

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

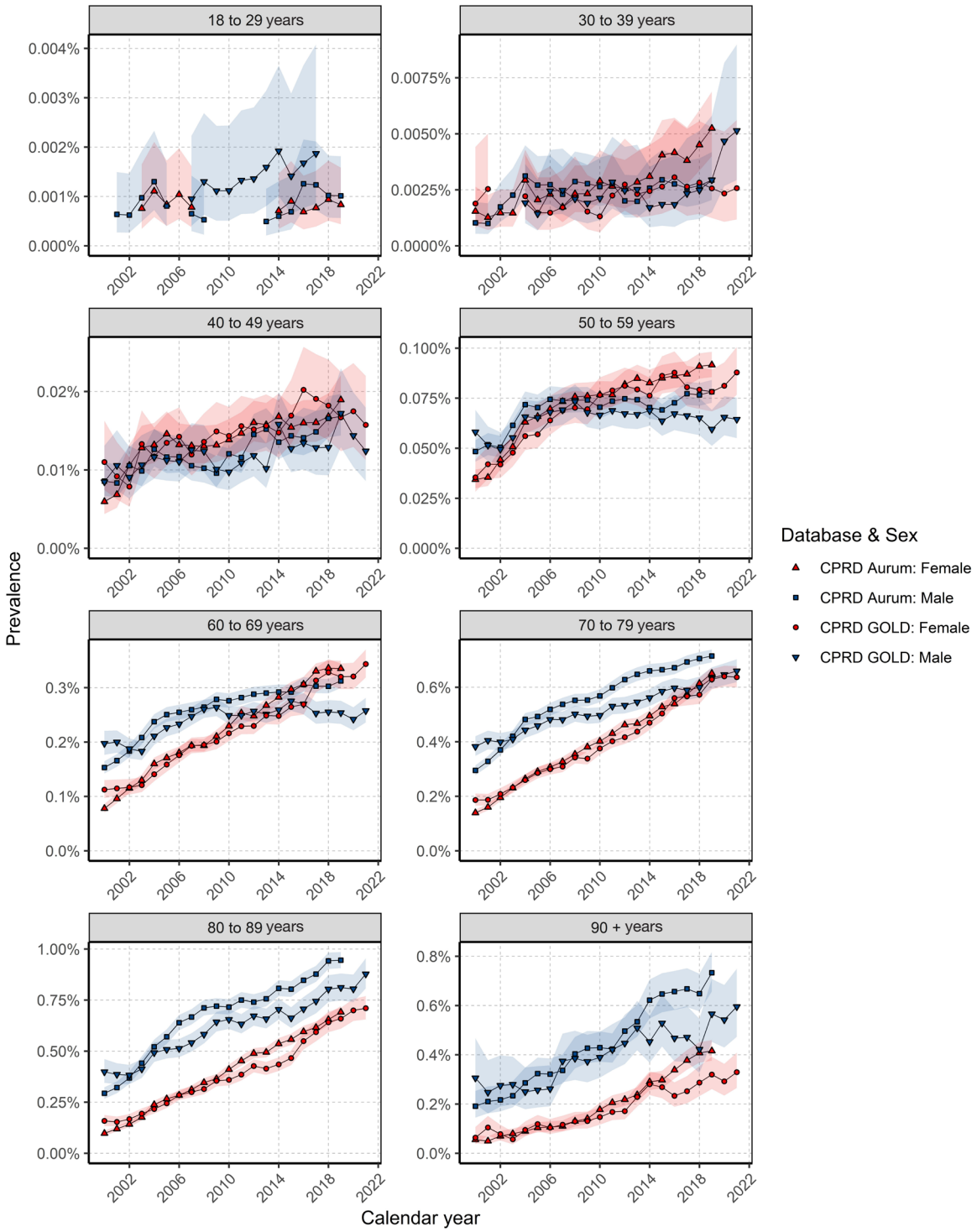

Figure S5 Annualised prevalence for lung cancer from 2000 to 2021 stratified by database, sex and age group. Band represents 95% CI.

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

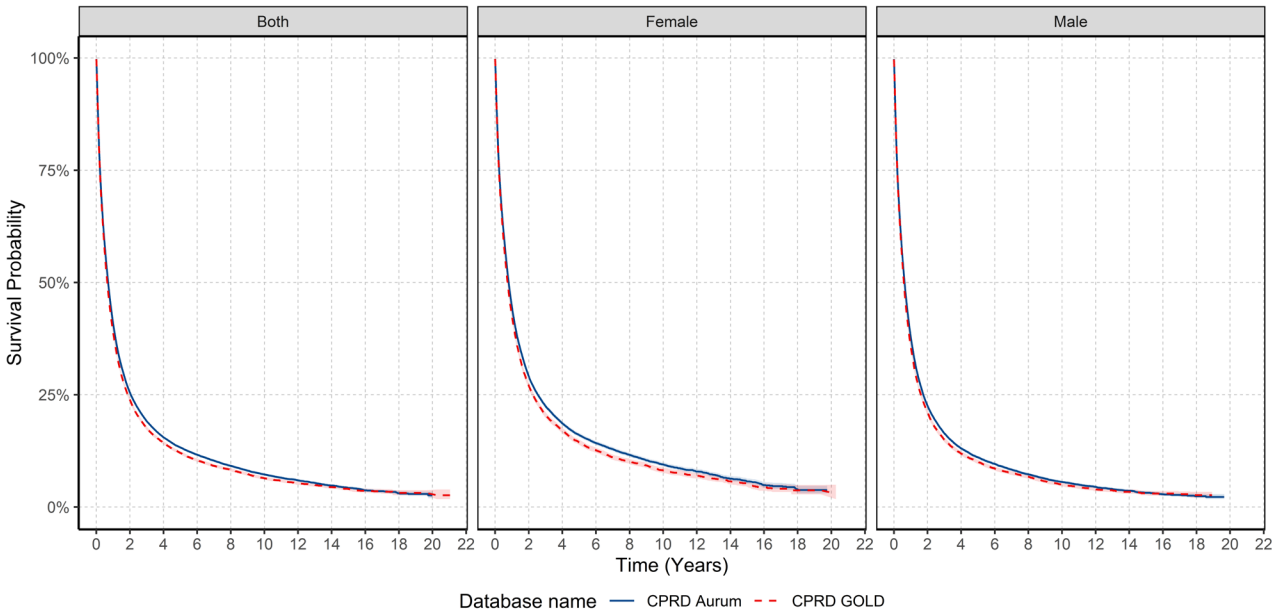

Figure S6 Kaplan-Meier survival curve of lung cancer by database and sex.

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

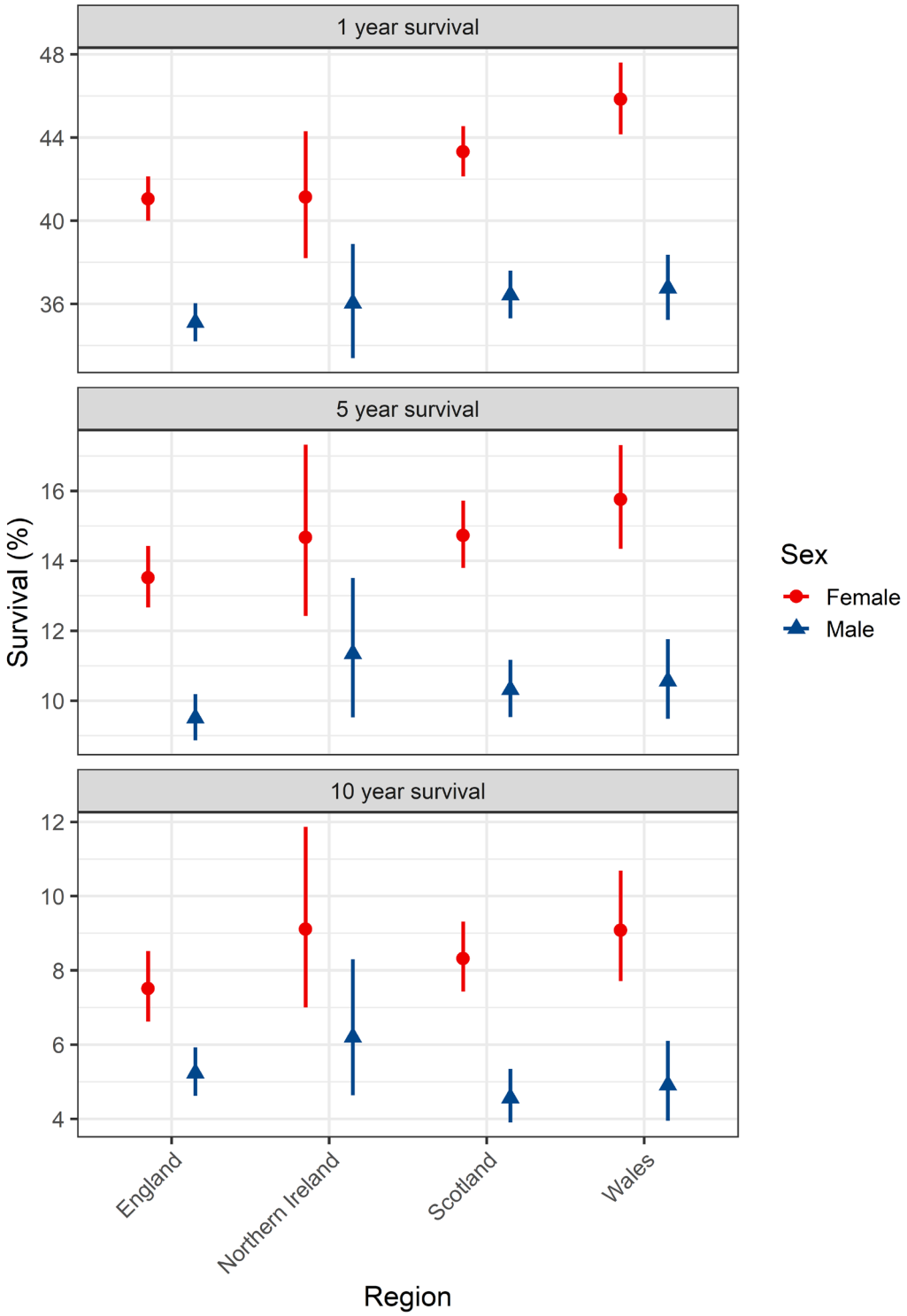

Figure S7 Survival (%) after 1, 5 and 10 years after lung cancer diagnosis stratified by UK region and sex for GOLD.

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