Risk of lung cancer increases with smoking in an English CPRD cohort

DASC503: Using Routine Data for Public Health

Risk of lung cancer increases with smoking in an English CPRD cohort

Name	Value
Date	31/12/2024
Word Count	1550
Student ID	201857199

Generative Artificial Intelligence (GAI)

I **did not** use GAI in the preparation of this work

Contents


```
library(multidplyr)
library(tidyverse)
library(readxl)
library(lubridate)
library(magrittr)
library(cowplot)
library(ggh4x)
library(epitools)
library(santoku)
```

```r
library(patchwork)
library(knitr)
library(kableExtra)
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

report::report(sessionInfo())
```

```
## Analyses were conducted using the R Statistical language (version 4.4.0; R Core
## Team, 2024) on Windows 11 x64 (build 22631), using the packages epitools
## (version 0.5.10.1; Aragon T, 2020), magrittr (version 2.0.3; Bache S, Wickham
## H, 2022), lubridate (version 1.9.3; Grolemund G, Wickham H, 2011), santoku
## (version 1.0.0; Hugh-Jones D, 2024), tibble (version 3.2.1; Müller K, Wickham
## H, 2023), patchwork (version 1.3.0; Pedersen T, 2024), ggh4x (version 0.2.8;
## van den Brand T, 2024), ggplot2 (version 3.5.1; Wickham H, 2016), forcats
## (version 1.0.0; Wickham H, 2023), multidplyr (version 0.1.3; Wickham H, 2023),
## stringr (version 1.5.1; Wickham H, 2023), tidyverse (version 2.0.0; Wickham H
## et al., 2019), readxl (version 1.4.3; Wickham H, Bryan J, 2023), dplyr (version
## 1.1.4; Wickham H et al., 2023), purrr (version 1.0.2; Wickham H, Henry L,
## 2023), readr (version 2.1.5; Wickham H et al., 2024), tidyr (version 1.3.1;
## Wickham H et al., 2024), cowplot (version 1.1.3; Wilke C, 2024), knitr (version
## 1.49; Xie Y, 2024) and kableExtra (version 1.4.0; Zhu H, 2024).
##
## References
## ----------
##   - Aragon T (2020). _epitools: Epidemiology Tools_. R package version 0.5-10.1,
## <https://CRAN.R-project.org/package=epitools>.
##   - Bache S, Wickham H (2022). _magrittr: A Forward-Pipe Operator for R_. R
## package version 2.0.3, <https://CRAN.R-project.org/package=magrittr>.
##   - Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate."
## _Journal of Statistical Software_, *40*(3), 1-25.
## <https://www.jstatsoft.org/v40/i03/>.
##   - Hugh-Jones D (2024). _santoku: A Versatile Cutting Tool_. R package version
## 1.0.0, <https://CRAN.R-project.org/package=santoku>.
##   - Müller K, Wickham H (2023). _tibble: Simple Data Frames_. R package version
## 3.2.1, <https://CRAN.R-project.org/package=tibble>.
##   - Pedersen T (2024). _patchwork: The Composer of Plots_. R package version
## 1.3.0, <https://CRAN.R-project.org/package=patchwork>.
##   - R Core Team (2024). _R: A Language and Environment for Statistical
## Computing_. R Foundation for Statistical Computing, Vienna, Austria.
## <https://www.R-project.org/>.
##   - van den Brand T (2024). _ggh4x: Hacks for 'ggplot2'_. R package version
## 0.2.8, <https://CRAN.R-project.org/package=ggh4x>.
##   - Wickham H (2016). _ggplot2: Elegant Graphics for Data Analysis_.
## Springer-Verlag New York. ISBN 978-3-319-24277-4,
## <https://ggplot2.tidyverse.org>.
##   - Wickham H (2023). _forcats: Tools for Working with Categorical Variables
## (Factors)_. R package version 1.0.0,
## <https://CRAN.R-project.org/package=forcats>.
##   - Wickham H (2023). _multidplyr: A Multi-Process 'dplyr' Backend_. R package
## version 0.1.3, <https://CRAN.R-project.org/package=multidplyr>.
##   - Wickham H (2023). _stringr: Simple, Consistent Wrappers for Common String
## Operations_. R package version 1.5.1,
```

<https://CRAN.R-project.org/package=stringr>.
- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G,
Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K,
Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K,
Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_,
4(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.
- Wickham H, Bryan J (2023). _readxl: Read Excel Files_. R package version
1.4.3, <https://CRAN.R-project.org/package=readxl>.
- Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar
of Data Manipulation_. R package version 1.1.4,
<https://CRAN.R-project.org/package=dplyr>.
- Wickham H, Henry L (2023). _purrr: Functional Programming Tools_. R package
version 1.0.2, <https://CRAN.R-project.org/package=purrr>.
- Wickham H, Hester J, Bryan J (2024). _readr: Read Rectangular Text Data_. R
package version 2.1.5, <https://CRAN.R-project.org/package=readr>.
- Wickham H, Vaughan D, Girlich M (2024). _tidyr: Tidy Messy Data_. R package
version 1.3.1, <https://CRAN.R-project.org/package=tidyr>.
- Wilke C (2024). _cowplot: Streamlined Plot Theme and Plot Annotations for
'ggplot2'_. R package version 1.1.3,
<https://CRAN.R-project.org/package=cowplot>.
- Xie Y (2024). _knitr: A General-Purpose Package for Dynamic Report Generation
in R_. R package version 1.49, <https://yihui.org/knitr/>. Xie Y (2015).
Dynamic Documents with R and knitr, 2nd edition. Chapman and Hall/CRC, Boca
Raton, Florida. ISBN 978-1498716963, <https://yihui.org/knitr/>. Xie Y (2014).
"knitr: A Comprehensive Tool for Reproducible Research in R." In Stodden V,
Leisch F, Peng RD (eds.), _Implementing Reproducible Computational Research_.
Chapman and Hall/CRC. ISBN 978-1466561595.
- Zhu H (2024). _kableExtra: Construct Complex Table with 'kable' and Pipe
Syntax_. R package version 1.4.0,
<https://CRAN.R-project.org/package=kableExtra>.

1 Introduction

1.1 Background

Lung cancer is a heterogeneous and common group of bronchogenic carcinomas presenting in the lower respiratory tract[1–4]. Risk factors include sex, smoking, air pollution, exposure to asbestos or other particulate metals, polycyclic hydrocarbons, co-morbidity with diseases such as idiopathic pulmonary fibrosis, genetic predisposition and socioeconomic deprivation/urbanisation[5–9]. The national increase in population levels, shift in demographic trends and temporal morphing of rates of exposure/prevalence of risk factors hence present an extant need to understand current and future burdens of lung cancer to optimise methodology, deployment and resource allocation for screening, diagnosis and therapy to promote health equity[10].

In this study, we examine age, sex, region, and deprivation specific incidence rates as well as additional risk of lung cancer due to smoking in a synthetic, closed CPRD cohort by integrating patient-level data from 2013-2020 with demographic attributes[11].

1.2 Cohort definition

Inclusion criteria: A retrospective closed cohort was constructed from routinely collected, anonymised primary care data in England through the auspices of the CPRD's Aurum dataset[11]. Participants were considered for the study if they had atleast one year of prior monitoring available on the CPRD dataset

(between 2013-2014); their data met CPRD's research quality standards; and they were aged between 25 and 90 at any point during the study period. Relative risk calculations were conducted based on the patients voluntary (and self-reported) exposure to smoking and incidence of disease.

```
dateFrom = ymd_hms("2014-1-1 0:0:1")
dateTo = ymd_hms("2020-12-31 23:59:59")

intervalStudy <- interval(dateFrom,dateTo)

intAgeMin = 25
intAgeMax = 90
```

Case definition: The event of interest for calculating incidence was defined as a diagnosis of cancer identified through the first assignment of a previously published SNOMED codelist[12]. For relative risk calculations, patient-reported exposure was defined as the patient voluntarily smoking (SNOMED 77176002), had previously voluntarily smoked (ex-smoker; SNOMED 8517006 or 405746006) or had not smoked at all (non-smoker; SNOMED 266919005 or 8392000)[13]. No data on dose or duration of smoking was available.

```
# Patient population data
tblData <- read_csv("Source/pop_snomed_assignment.csv")
# SNOMED codelist
tblLungCancerCodes <- read_csv("Source/dLungCancerCodes.csv",
    col_types = cols(medcodeid = col_character(),
        snomedctconceptid = col_character(),
        snomedctdescriptionid = col_character()))
tblLungCancerCodes %>% select(descr,snomedctconceptid,snomedctdescriptionid)
```

```
## # A tibble: 38 x 3
##    descr                                snomedctconceptid snomedctdescriptionid
##    <chr>                                <chr>             <chr>
##  1 Adenocarcinoma of lung               254626006         379172013
##  2 Non-small cell lung cancer           254637007         379195016
##  3 Primary malignant neoplasm of lung   93880001          510792012
##  4 Referral by lung cancer nurse specia~ 1863331000006107 1863331000006111
##  5 Small cell lung cancer               254632001         379181019
##  6 Squamous cell carcinoma of lung      254634000         379184010
##  7 [RFC] Lung cancer                    907111000006102   907111000006118
##  8 [x]malignant neoplasm of bronchus or~ 363358000        396221000006112
##  9 Malignant neoplasm of trachea, bronc~ 430621000        2765453013
## 10 Malignant neoplasm of trachea        363432004         482662016
## # i 28 more rows
```

```
#tblLungCancerCodes <- read_csv("https://raw.githubusercontent.com/annalhead/
#CPRD_multimorbidity_codelists/9d26739d93744c8444aedbe10de65657c4af6bc0/
#codelists/Primary%20Malignancy_Lung.csv")
```

Exclusion and exit criteria: A patient was excluded from the study if they had a confirmed diagnosis of lung cancer before the 1st of January 2014 or they died from lung cancer without a confirmed diagnosis in the study period. Patients were censored at the earliest date on which they were either diagnosed with lung cancer, died from any cause or reached the end of the study defined as the end of 31st of December 2020.

2 Methods

Overall crude incidence rates per 10,000 person-years at risk were calculated, and results were stratified by 5 year age bands, sex, and finally either index of multiple deprivation or region (defined by the strategic health authority responsible for their care) as the case maybe. Age-specific rates were calculated from fractional patient year calculations per age band on a per patient basis. This approach takes into account their date of birth, the study duration, their date of censoring. For example, if a patient participated for more than a year, their patient year contributions per age would be considered (eg. a patient would contribute 0.7 patient years when they were 33 years of age, 1 for 34, and 0.6 patient years for 35 years of age).

2.1 (Assessments 1 & 2) Helper functions

Incase of non-unique values for an individual in any variable, we assume the most frequently occurring value as the true value. For instance, if a patient who participated in 5 years of the study was an ex-smoker for all but one year (during which they smoked), the patient will be regarded as an ex-smoker. However if there is a record of a patient ever having smoked, they will always be regarded as an ex-smoker (eg. patient 661552) unless they are a current smoker. This assumption however incorrectly assumes that even a single incidence of smoking will permanently alter the risk of lung cancer, and future studies must take into account the duration and dose of smoking to calculate specific risks as well as relations between proximity of an exposure to incidence of cancer.

```
# Function to get the most frequently occurring value in a vector.
fGetMostFrequent <- function(tblInput) {
  tblInput %>%
    table() %>%
    sort(decreasing = TRUE) %>%
    .[1] %>%
    names() -> output
  return(output)
}

# Function to get the most frequently occurring value in a vector. Specific for
# smokers as if a person has ever smoked, they will be an ex-smoker. This
# however has limitations in that if a person smokes for only 1 months they will
# be considered equally at risk as someone who has previously smoked for 4 years.

fGetMostFrequentSmoker <- function(tblInput) {
  # Determine the most frequent classification
  tblInput %>%
    table() %>%
    sort(decreasing = TRUE) -> Temp

  Temp[1] %>%
    names() -> output

  # Check if there's any "smoker" or "ex-smoker" in the input
  # which is not the most frequent
  if (any(c("smoker", "ex_smoker") %in% names(Temp)) &&
      !(output %in% c("smoker", "ex-smoker"))) {
    return("ex_smoker")
  }
  else
    return(output) # Otherwise, return the most frequent classification
```

```
}

# Function to calculate the time difference in years between two dates
fCalcYear <- function(dateFrom, dateTo) {
  return(interval(dateFrom, dateTo) / years(1))
}

# Function to save ggPlot
fSavePlot <-
  function(plot,
           title = NA,
           subtitle = NA,
           h = 8,
           w = 12,
           background = "white") {
    ggsave(
      plot,
      filename = paste0(title, subtitle, ".png"),
      path = "Figures/",
      device = "png",
      dpi = 800,
      width = w,
      height = h,
      units = "in",
      bg = background,
      create.dir = TRUE
    )
  }

# Function to calculate incidence and poission CI
fCalculateIncidence <-
  function(tblNumerator,
           tblDenominator,
           strGroupVars)
  {
    tblNumerator %>%
      group_by(across(all_of(strGroupVars))) %>%
      summarise(diagnosed = sum(ever_diagnosed)) %>%
      left_join(tblDenominator %>%
                  group_by(across(all_of(strGroupVars))) %>%
                  summarise(patient_years = sum(total_patient_years))) %>%
      mutate(
        Incidence = diagnosed / patient_years * 10000,
        CI = pois.approx(diagnosed, pt = patient_years) * 10000
      )
  }

# Function to plot incidence and poission CI
fPlotIncidence <-
  function(tblData,
           strTitle,
           strGroupBy,
           strNameGroupBy,
```

```
          numCrudeGlobalIncidence,
          strSmooth,
          lstYLim = c(0, 100)) {
    tblData %>%
      ggplot(aes(
        x = ageband,
        y = Incidence,
        group = !!sym(strGroupBy),
        colour = !!sym(strGroupBy)
      )) +
      scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
      geom_hline(aes(yintercept = numCrudeGlobalIncidence)) +
      geom_errorbar(aes(ymin = CI$lower, ymax = CI$upper),
                    width = 0.05,
                    alpha = 0.2) +
      geom_point() +
      geom_smooth(method = strSmooth, se = FALSE) +
      ggsci::scale_color_aaas() +
      theme_cowplot(12) +
      ylab("Incidence rate per 10k person years") +
      xlab("Age") +
      coord_cartesian(ylim = lstYLim) +
      labs(colour = strNameGroupBy, title = strTitle)
  }

fPlotBoxIncidence <-
  function(tblData,
           strX,
           strY,
           strTitle,
           strGroupBy,
           boolStaggerXLbl = TRUE,
           strSmooth,
           numCrudeGlobalIncidence,
           lstYLim = c(0, 100)) {
    tblData %>%
      ggplot(aes(
        x = !!sym(strX),
        y = !!sym(strY),
        group = !!sym(strGroupBy),
        colour = !!sym(strGroupBy)
      )) +
      {
        if (boolStaggerXLbl)
          scale_x_discrete(guide = guide_axis(n.dodge = 3))
      } +
      geom_hline(aes(yintercept = numCrudeGlobalIncidence)) +
      geom_errorbar(
        aes(ymin = CI$lower, ymax = CI$upper),
        width = 0.05,
        alpha = 0.2,
        show.legend = FALSE
      ) +
```

```
    geom_boxplot(show.legend = FALSE) +
    ggsci::scale_color_aaas() +
    theme_cowplot(12) +
    ylab("Incidence rate per 10k person years") +
    xlab("") +
    coord_cartesian(ylim = lstYLim) +
    labs(title = strTitle)
}
```

2.2 (Assessments 1 & 2) Data cleaning and transformation

2.2.1 Data restructure

```
tblData %<>%
  # Remove rows with missing SNOMED description IDs
  filter(!is.na(snomedctdescriptionid)) %>%
  mutate(
    # Convert date columns to Date objects
    event_date = ymd(event_date),
    dob = ymd(dob),

    # Identify cancer-relevant events
    is_cancer_relevant = ifelse(
      snomedctdescriptionid %in% tblLungCancerCodes$snomedctdescriptionid |
        medcodeid %in% tblLungCancerCodes$medcodeid,
      TRUE,
      FALSE
    ),

    # Identify death events
    is_death = ifelse(snomedctdescriptionid < 0, TRUE, FALSE),
    is_death_cancer = ifelse(snomedctdescriptionid == -5, TRUE, FALSE),

    # Determine the censor date
    censor_date = case_when(
      is_cancer_relevant == TRUE ~ event_date,
      is_death == TRUE ~ event_date,
      .default = dateTo # Default censor date if no event = end of study
    )
  )
```

2.2.2 Parallel aggregation with `multidplyr`

```
# Create a parallel processing cluster using available CPU cores
pllCluster <- new_cluster(parallel::detectCores() - 4)

# Distribute data across the cluster by patient ID
tblData %>%
  group_by(pid) %>%
```

```r
  partition(pllCluster) -> pllDataSummary

# Load required libraries and functions on all cluster nodes
cluster_library(pllCluster, "tidyverse")
cluster_copy(pllCluster, c("fCalcYear",
                           "fGetMostFrequent",
                           "fGetMostFrequentSmoker",
                           "dateFrom"))

# Summarise data per patient across cluster nodes
pllDataSummary %>%
  summarise(
    # If non-unique values exist, get the most frequent values for
    sex = fGetMostFrequent(sex),
    qimd = fGetMostFrequent(qimd),
    ethnicity = fGetMostFrequent(ethnicity),
    sha = fGetMostFrequent(sha),
    smoking_status = fGetMostFrequentSmoker(smoking_status),
    dob = fGetMostFrequent(dob),

    # Check if ever diagnosed with cancer
    ever_diagnosed = max(is_cancer_relevant),

    # Check if patient ever died
    ever_died = max(is_death),

    # Check if death was cancer-related
    ever_died_cancer = max(is_death_cancer),

    # Get earliest date of event for a patient.
    censor_date = min(censor_date)
  ) %>%

  # Calculate patient age at event date, rounded down.
  # We need this for age-specific incidence calculations
  mutate(age = fCalcYear(dob, censor_date) %>% floor()) %>%

  # Bring results back to the main R session
  collect() %>%

  # Remove patients who die of cancer, but are not diagnosed with cancer within
  # available data. These patients would have likely contracted the disease
  # earlier on, and as they already have the disease, they are not at risk.
  # Also only keep events within study interval.
  filter(!(ever_diagnosed == 0 & ever_died_cancer == 1),
         censor_date %within% intervalStudy) %>%

  # For some reason, lubridate breaks down when adding years to a leap year..
  # possible bug? Anyway, this should round down the leap day to the previous
  # day. Not an ideal solution, but given the average person-year length of the
  # study, the contribution of a day shouldn't add up that much.
  mutate(dob = gsub("02-29","02-28", dob)) -> tblDataSummary
```

2.3 (Assessment 1) Numerator calculations

```
# Filter patients diagnosed with cancer based on predefined
# cutoffs and discretise ages
tblDataSummary %>%
  filter(ever_diagnosed == 1,
         (age >= intAgeMin &
            age <= intAgeMax)) %>%
  mutate(ageband = chop_width(
    age,
    start = 25,
    width = 5,
    labels = lbl_discrete()
  )) -> tblNumerator

# Save the filtered dataset as a CSV file
tblNumerator %>% write_csv("tblNumerator.csv")

# Creating an empty tibble with no unique diagnoses of cancer to capture those
# years, sex, qimd, etc in which there are no diagnoses of cancer (incidence
# has been 0). As such, these data are filtered out in our numerator
# calculations. This will introduce it back in.
# This is so that we can plot those years with no incidences easily
expand_grid(
  sex = unique(tblNumerator$sex) %>% as.character(),
  sha = unique(tblNumerator$sha) %>% as.character(),
  qimd = unique(tblNumerator$qimd) %>% as.character(),
  ethnicity = unique(tblNumerator$ethnicity) %>% as.character(),
  smoking_status = unique(tblNumerator$smoking_status) %>% as.character(),
  ageband = santoku::chop_width(
    seq(intAgeMin, intAgeMax, by = 5),
    start = 25,
    width = 5,
    labels = lbl_discrete()
  )
) %>%
  mutate(
    pid = 0,
    ever_diagnosed = 0,
    ever_died = 0,
    ever_died_cancer = 0,
    dob = NA,
    censor_date = NA,
    age = NA,
  ) %>%
  bind_rows(tblNumerator) -> tblGraphNumerator
```

2.4 (Assessment 1) Denominator calculations

```
# WARNING: This is computationally expensive. The need to calculate
# each person's person-year calculation, per age creates a tibble with 3m rows.
```

```r
# This can be done more efficiently, for instance by not
# calculating the py for every year, but merely the first year in the study and
# the year of censoring. I have run this analysis and saved the results as a csv
# file, which will be used further on, but have included the code for clarity.

# Create a parallel processing cluster using available CPU cores
pllCluster <- new_cluster(parallel::detectCores() - 4)

# Distribute data across the cluster by patient ID
tblDataSummary %>%
group_by(pid) %>%
partition(pllCluster) -> pllDataSummary

# Load required libraries and functions on all cluster nodes
cluster_library(pllCluster, "tidyverse")
cluster_copy(pllCluster, c("fCalcYear", "dateFrom"))

# We are now going to achieve two objectives with the next section. We are going
# to calculate all ages a person experienced in the study, and the total
# person- year they contributed to the study after taking into account their dob

# We initially calculate age at study start and censor date. Then we will create
# a new row for each age the patient experienced in the study. Eg. if a patient
# born on 2000 participates in the study from 2014 till 2018 then the patient
# will have been 14,15,16,17, and 18. This is important for age specific person
# year calculations as they will have contributed person years over many ageband.

pllDataSummary %>%
mutate(
age_at_start = fCalcYear(dob, dateFrom) %>% floor(),
age_at_censor = fCalcYear(dob, censor_date) %>% floor(),
age_seq = map2(age_at_start, age_at_censor, seq)
) %>%
collect() -> Temp

# unnesting not implemented in multidplyr. So we have to exit out of multidplyr
# and reparallelise.
Temp %<>%
unnest_longer(age_seq)

# While we have calculated the ages the patient contributed, we have to
# calculate the fractional person-years per age they took part in the study.
# Eg. a patient born on 2000-01-01 vs 2000-06-01 will have contributed different
# person years (by 6 months) at any point in the study. This approach takes
# their dob into account and calculates the correct factional person-years.
Temp %<>%
group_by(pid) %>%
partition(pllCluster) %>%
mutate(
year_start = pmax(dateFrom, ymd(dob) + years(age_seq)),
year_end = pmin(censor_date, ymd(dob) + years(age_seq + 1)),
patient_years = fCalcYear(year_start, year_end)
) %>%
```

```
collect()

# For the cases where the patient's birthday is after the censor date.
# The fraction of the year up to the censor date is correctly calculated in
# the previous row. However, the loop may generate an additional row for the
# next birthday, resulting in a negative 'patient_years' value because the
# interval extends past the censor date. This filter ensures only valid rows
# are kept by removing any where the calculated interval is invalid.

# Group by demographic and age attributes, then calculate total py per group
Temp %>%
filter(patient_years >= 0) %>%
group_by(sex, qimd, ethnicity, sha, smoking_status, age = age_seq) %>%
summarise(total_patient_years = sum(patient_years, na.rm = TRUE)) %>%
arrange(age) %>%

# Filtering here as we do not want to include at-risk times of people who
# are outside of the age cut-offs
filter(age >= intAgeMin &
age <= intAgeMax) %>%
mutate(ageband = chop_width(
age,
start = 25,
width = 5,
labels = lbl_discrete()
))  -> tblDenominator

# Save the final aggregated denominator table to a CSV file
tblDenominator %>% write_csv("tblDenominator.csv")

# Clean variables
rm(Temp, pllCluster, pllDataSummary, tblData, tblLungCancerCodes)
```

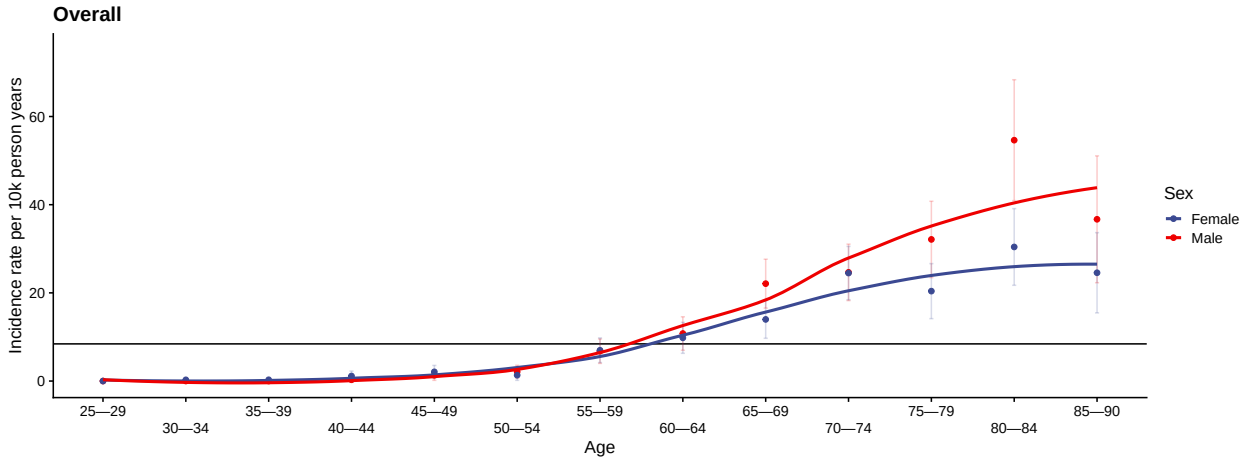
```
#To save myself recalculating the denominator table everytime I modify the Rmd,
# I save the output to a csv and read it in.

tblDenominator <- read_csv("tblDenominator.csv")

# Calculate unstratified national incidence for comparison
numCrudeGlobalIncidence = sum(tblNumerator$ever_diagnosed)/
  sum(tblDenominator$total_patient_years) * 10000
```

3 Results

3.1 (Assessment 1) Age and sex specific incidence rate ratios for lung cancer are higher for men than women

```
lstGroupVar = c("sex", "ageband")
ggVarGroupBy = "sex"
ggNameGroupBy = "Sex"
```

```
strGraphTitle = "SexAge"
strSmoother = "loess"


fCalculateIncidence(tblGraphNumerator,
                    tblDenominator,
                    lstGroupVar) %>%
   mutate(sex = gsub("^men","Male", sex),
          sex = gsub("^women","Female", sex)) -> tblTemp


tblTemp %>% fPlotIncidence(
  strTitle = "Overall",
  strGroupBy = ggVarGroupBy,
  strNameGroupBy = ggNameGroupBy,
  strSmooth = strSmoother,
  numCrudeGlobalIncidence = numCrudeGlobalIncidence,
  lstYLim = c(0,75)
) -> ggTempAll

ggTempAll %>% fSavePlot(strGraphTitle,strSmoother, w = 6, h = 6)

ggTempAll
```




```
fCalculateIncidence(tblGraphNumerator,
                    tblDenominator,
                    ggVarGroupBy) %>%
  mutate(sex = gsub("^men", "Male", sex),
         sex = gsub("^women", "Female", sex)) %>%
  mutate(order = ifelse(sex == "Female", 1,2)) %>%
  arrange(order) %>%
  column_to_rownames(ggVarGroupBy) %>%
  select(diagnosed, patient_years) %>%
  as.matrix() %>%
  rateratio() %>%
  kable()
```

	diagnosed	patient_years
Female	293	377025.8
Male	325	356265.3
Total	618	733291.1

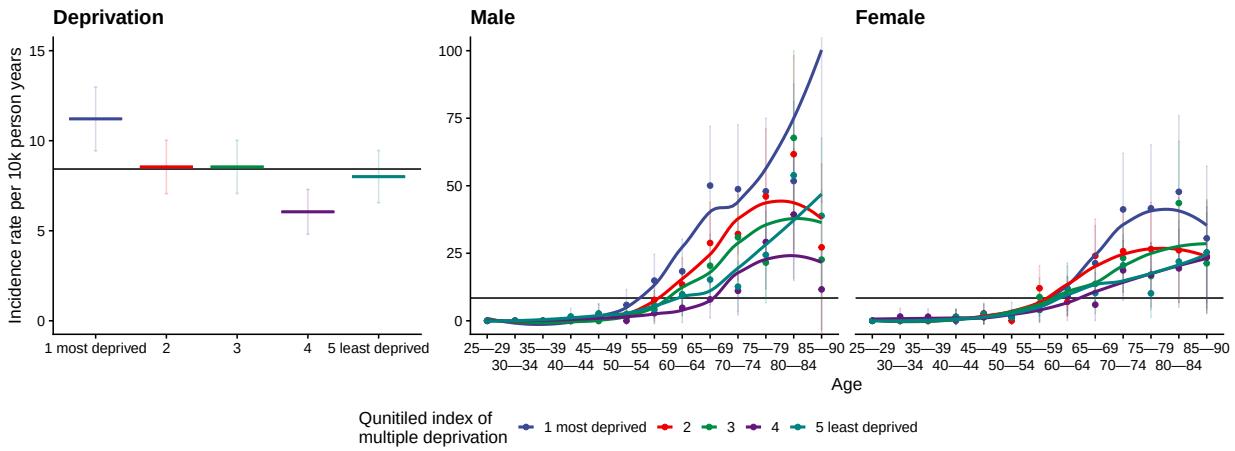
	estimate	lower	upper
Female	1.000000	NA	NA
Male	1.173786	1.002425	1.37519

	midp.exact	wald
Female	NA	NA
Male	0.046602	0.0463887

3.2 (Assessment 1) Age, sex and deprivation specific incidence rates for lung cancer are directly proportional to quintiled index of multiple deprivation

```
lstGroupVar = c("qimd", "ageband")
ggVarGroupBy = "qimd"
ggNameGroupBy = "Qunitiled index of \nmultiple deprivation"
strGraphTitle = "AgeQIMD"
strSmoother = "loess"


fCalculateIncidence(
  tblGraphNumerator %>% filter(sex == "men"),
  tblDenominator %>% filter(sex == "men"),
  lstGroupVar
) -> tblTempMale

fCalculateIncidence(
  tblGraphNumerator %>% filter(sex == "women"),
  tblDenominator %>% filter(sex == "women"),
  lstGroupVar
) -> tblTempFemale

tblTempMale %>% fPlotIncidence(
  strTitle = "Male",
  strGroupBy = ggVarGroupBy,
  strNameGroupBy = ggNameGroupBy,
  strSmooth = strSmoother,
  numCrudeGlobalIncidence = numCrudeGlobalIncidence
) -> ggTempMale

tblTempFemale %>% fPlotIncidence(
  strTitle = "Female",
  strGroupBy = ggVarGroupBy,
  strNameGroupBy = ggNameGroupBy,
  strSmooth = strSmoother,
  numCrudeGlobalIncidence = numCrudeGlobalIncidence
) -> ggTempFemale

fCalculateIncidence(tblNumerator, tblDenominator, c("qimd")) %>%
    fPlotBoxIncidence(strX = "qimd",
                      strY = "Incidence",
                      strTitle = "Deprivation",
                      strGroupBy = "qimd",
                      boolStaggerXLbl = FALSE,
```

```
                    numCrudeGlobalIncidence =  numCrudeGlobalIncidence,
                    lstYLim =  c(0,15)) -> ggTemp

ggFig <- ggTemp +
  ggTempMale +
  ggTempFemale +
  plot_layout(axes = "collect", axis_titles = "collect", guides = "collect") &
  theme(legend.position = 'bottom')

ggFig %>% fSavePlot(strGraphTitle,strSmoother, w = 14, h = 6)
ggFig
```




```
fCalculateIncidence(tblGraphNumerator,
                    tblDenominator,
                    ggVarGroupBy) %>%
  mutate(order = ifelse(qimd == "5 least deprived", 1,2)) %>%
  arrange(order) %>%
  column_to_rownames(ggVarGroupBy) %>%
  select(diagnosed, patient_years) %>%
  as.matrix() %>%
  round(2) %>%
  rateratio() %>%
  kable()
```

	diagnosed	patient_years
5 least deprived	117	146149.4
1 most deprived	154	137313.1
2	127	148593.6
3	129	150916.9
4	91	150318.2
Total	618	733291.1

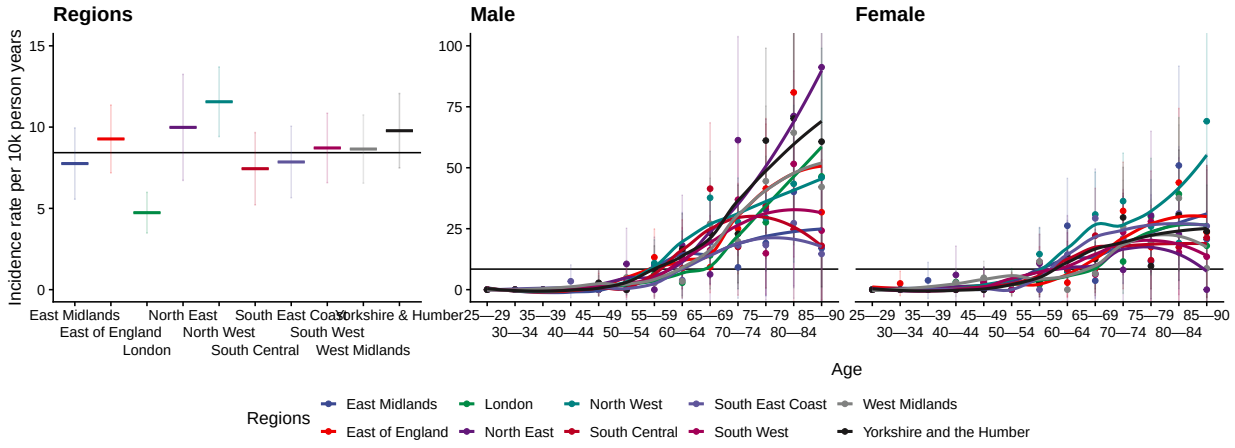
	estimate	lower	upper
5 least deprived	1.0000000	NA	NA
1 most deprived	1.4003990	1.1021804	1.7843425
2	1.0674959	0.8303190	1.3738255
3	1.0675934	0.8312973	1.3726222
4	0.7565154	0.5738836	0.9941288

	midp.exact	wald
5 least deprived	NA	NA
1 most deprived	0.0058075	0.0057430
2	0.6104979	0.6095849
3	0.6086319	0.6076385
4	0.0452419	0.0448805

3.3 (Assessment 1) Age, sex and region specific incidence rates are higher in the north of England

```
lstGroupVar = c("sha", "ageband")
ggVarGroupBy = "sha"
ggNameGroupBy = "Regions"
strGraphTitle = "AgeSha"
strSmoother = "loess"


fCalculateIncidence(
  tblGraphNumerator %>% filter(sex == "men"),
  tblDenominator %>% filter(sex == "men"),
  lstGroupVar
) -> tblTempMale

fCalculateIncidence(
  tblGraphNumerator %>% filter(sex == "women"),
  tblDenominator %>% filter(sex == "women"),
  lstGroupVar
) -> tblTempFemale

fCalculateIncidence(tblGraphNumerator,
                    tblDenominator,
                    lstGroupVar) -> tblTemp


tblTempMale %>% fPlotIncidence(
  strTitle = "Male",
  strGroupBy = ggVarGroupBy,
  strNameGroupBy = ggNameGroupBy,
  strSmooth = strSmoother,
  numCrudeGlobalIncidence = numCrudeGlobalIncidence
) -> ggTempMale

tblTempFemale %>% fPlotIncidence(
  strTitle = "Female",
  strGroupBy = ggVarGroupBy,
```

```
  strNameGroupBy = ggNameGroupBy,
  strSmooth = strSmoother,
  numCrudeGlobalIncidence = numCrudeGlobalIncidence
) -> ggTempFemale

fCalculateIncidence(tblNumerator, tblDenominator, c("sha")) %>%
   mutate(sha =
            gsub("Yorkshire and the Humber", "Yorkshire \\& Humber", sha)) %>%
  fPlotBoxIncidence(strX = "sha",
                    strY = "Incidence",
                    strTitle = "Regions",
                    strGroupBy = "sha",
                    boolStaggerXLbl = TRUE,
                    numCrudeGlobalIncidence =  numCrudeGlobalIncidence ,
                    lstYLim =  c(0,15)) -> ggTemp

ggFig <- ggTemp +
  ggTempMale +
  ggTempFemale +
  plot_layout(axes = "collect", axis_titles = "collect", guides = "collect") &
  theme(legend.position = 'bottom')

ggFig %>% fSavePlot(strGraphTitle,strSmoother, w = 14, h = 6)
ggFig
```




```
fCalculateIncidence(tblGraphNumerator,
                    tblDenominator,
                    ggVarGroupBy) %>%
  mutate(order = ifelse(sha == "London", 1,2)) %>%
  arrange(order) %>%
  column_to_rownames(ggVarGroupBy) %>%
  select(diagnosed, patient_years) %>%
  as.matrix() %>%
  rateratio() %>%
  kable()
```

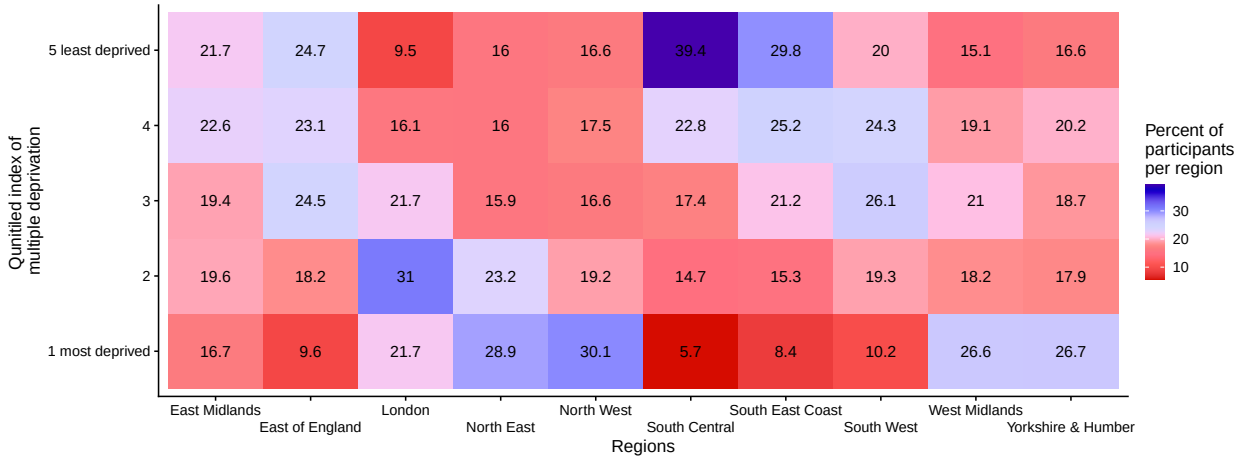
	diagnosed	patient_years
London	55	116162.90
East Midlands	48	61885.34
East of England	76	81969.72
North East	36	36060.99
North West	112	96884.86
South Central	43	57790.73
South East Coast	49	62379.14
South West	64	73418.98
West Midlands	65	75144.01
Yorkshire and the Humber	70	71594.46
Total	618	733291.12

	estimate	lower	upper
London	1.000000	NA	NA
East Midlands	1.638886	1.108494	2.414992
East of England	1.956552	1.385329	2.781635
North East	2.111979	1.374810	3.203772
North West	2.437930	1.773900	3.390786
South Central	1.572823	1.049410	2.341606
South East Coast	1.659668	1.125217	2.440786
South West	1.840219	1.283314	2.648041
West Midlands	1.826092	1.275298	2.624253
Yorkshire and the Humber	2.063679	1.450820	2.950338

	midp.exact	wald
London	NA	NA
East Midlands	0.0135533	0.0115933
East of England	0.0001348	0.0001097
North East	0.0007944	0.0003696
North West	0.0000000	0.0000000
South Central	0.0285431	0.0251194
South East Coast	0.0108654	0.0091959
South West	0.0009317	0.0007480
West Midlands	0.0010276	0.0008398
Yorkshire and the Humber	0.0000564	0.0000390

3.4 (Assessment 1) Relation between deprivation and region show more deprived regions are in the North of England

```
tblDataSummary %>%
  select(qimd, sha) %>%
  table() %>%
  as_tibble(rownames = "Region") %>%
  group_by(sha) %>%
  mutate(Percent = n/sum(n)*100,
         sha = gsub("Yorkshire and the Humber",
                    "Yorkshire \\& Humber", sha)) -> ggTemp

ggTemp %>%
  ggplot(aes(sha, qimd, fill = Percent)) +
  geom_tile() +
  geom_text(aes(label = round(Percent, 1))) +
   ggsci::scale_fill_gsea(reverse = TRUE) +
   theme_cowplot(12) +
   scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  ylab("Qunitiled index of \nmultiple deprivation") +
  xlab("Regions") +
  labs(fill = "Percent of \nparticipants \nper region")-> ggFig

ggFig %>% fSavePlot("RegionDeprivation","Htmp", w = 8, h = 6)
ggFig
```



3.5 (Assessment 2) Relative risks of lung cancer significantly increases with smoking

```
tblDataSummary %>%
  group_by(smoking_status, ever_diagnosed) %>%
  summarise(n = length(ever_diagnosed)) %>%
    mutate(smoking_status = gsub("^never_smoker","Non-smoker", smoking_status),
          smoking_status = gsub("^ex_smoker","Ex-smoker", smoking_status),
          smoking_status = gsub("^smoker","Current smoker", smoking_status),
          ever_diagnosed = gsub("^0","No lung cancer", ever_diagnosed),
          ever_diagnosed = gsub("^1","Has lung cancer", ever_diagnosed)) %>%
  pivot_wider(names_from = ever_diagnosed, values_from = n) %>%
  mutate(smoking_status = factor(smoking_status,
                                 levels = c("Non-smoker",
                                            "Ex-smoker",
                                            "Current smoker"))) %>%
  arrange(smoking_status) %>%
  column_to_rownames("smoking_status") %>%
  as.matrix() %>%
  epitab(method = "riskratio") %>%
  .$tab %>%
  kable()
```

	No lung cancer	p0	Has lung cancer	p1	riskratio	lower	upper	p.value
Non-smoker	52233	0.9986807	69	0.0013193	1.000000	NA	NA	NA
Ex-smoker	36014	0.9899667	365	0.0100333	7.605212	5.881975	9.833305	0
Current smoker	21257	0.9914183	184	0.0085817	6.504920	4.934926	8.574393	0

Our analyses show that smoking significantly increases the risk of lung cancer, and quitting smoking does not immediately attenuate the risk. Indeed, with non-smokers as a reference, current smokers have a 550% (RR: 6.50 [95% CI: 4.93-8.57]; p <0.001) higher risk of contracting lung cancer whereas ex-smokers counterintutively display a 660% (RR: 7.60 [5.88-9.83]; p < 0.001) increased risk of lung cancer compared to non-smokers. The observed higher risk of lung cancer among ex-smokers compared to current smokers is often driven by survival biases, high latency periods (with cancers potentially requiring sustained exposure to smoking before development and once having been exposed, requiring a long time for the risk to reduce)[14], or reverse causation (with patients potentially quitting smoking upon development of symptoms)[15], rather than a genuine protective effect of continued smoking. Hence stratification by dose of exposure, age at smoking cessation, and time since cessation along with sensitivity analyses are essential for accurately assessing risks[16,17].

4 Critical appraisal

By using a relatively modest primary care dataset of 110,122 patients from CPRD Aurum, our study has been able to capture lung cancer incidence among a diverse sub-population in the UK. In line with other studies, we report that there is a region specific difference in incidence of lung cancer, with the North of England having higher disease burden[6,18]. We also report that smoking significantly increases the risk of lung cancer by a factor of 6-8[19,20]. While our study further supports such lung cancer trends, it remains limited by biases inherent in retrospective observational data and assumptions of static risk exposure.

While the CPRD datasets have been broadly suggested to be representative of the UK population[21,22], certain groups of people such as those who subscribe to private healthcare, those in prisons[23,24], are homeless[25,26], or in the armed forces[27] are underrepresented; all of whom have significantly different exposures to smoking potentially introducing participation biases.

A second significant limitation in the study is that every eligible patient was considered to be equally at risk for lung cancer (which is improbable[28]). Hence on a granular examination, there is a chance we may underestimate the true risk in an appropriatly risk-stratified sub-group.

As only participants from 25 years of age onwards were considered for the study due to constraints in data availability, they may however exclude rarer cancer events among younger adults[29], potentially decreasing the number of cancer diagnoses in the unexposed group and biasing risk calculations. Furthermore, these young adults may have non-monitored risk factors and differential survival trajectories than older patients with lung cancer, with important implications for health policy[30].

While the determination of smoking was implemented from a previously published code-list[13], there was no availability of dosage (eg. no of packs a day) or duration of previous exposure, both of which have a relevant impact on the risk profile. No data was also recorded on the effect of passive smoking by participants. Furthermore, it is also possible that such information may not be frequently updated within health records unless a diagnostic indication warrants recording such information[31,32]. This therefore may artificially enrich the proportion of exposed individuals incident with the disease in our risk calculations.

Further studies can be performed utilising mortality/survival analysis within a competing risk framework to link age based incidence and survival. This will give us insights into if older patients have worse survival outcomes, if regional or socioeconomic disparities affect survival (and if regional health policies align with these dispatities), how these findings align with incidence patterns, if transitioning from a smoker to an ex-smoker impacts disease incidence/survival.

5 Reflective summary

Through these assignments, I gained valuable skills in epidemiological data analysis, advanced R programming, and critical thinking. Perhaps the most important lesson learnt was that there is no one good answer/no one good method to anything, and every choice will introduce one or the other bias which we must work around. These skills were mostly learnt through hands on time with the assessments, and guided by discussions in lectures and practical classes. This learning matters because it strengthens my ability to conduct robust, reproducible analyses (and interpret such analyses performed by others) that can inform public health policies and interventions. Moving forward, I will apply these skills to future research projects, improving data handling, analysis efficiency, and communication of findings while maintaining a critical eye on potential biases and limitations.

References

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer* **136**, E359–E386 (2015).
2. Riaz, S. P. *et al.* Trends in incidence of small cell lung cancer and all lung cancer. *Lung cancer* **75**, 280–284 (2012).
3. Corby, G. *et al.* Incidence, prevalence, and survival of lung cancer in the united kingdom from 2000–2021: A population-based cohort study. *Translational Lung Cancer Research* **13**, (2024).
4. Travis, W. D. Pathology of lung cancer. *Clinics in chest medicine* **32**, 669–692 (2011).
5. Schabath, M. B. & Cote, M. L. Cancer progress and priorities: Lung cancer. *Cancer epidemiology, biomarkers & prevention* **28**, 1563–1579 (2019).

6. Riaz, S. P. *et al.* Lung cancer incidence and survival in england: An analysis by socioeconomic deprivation and urbanization. *Journal of Thoracic Oncology* **6**, 2005–2010 (2011).

7. Burns, D. M. Primary prevention, smoking, and smoking cessation: Implications for future trends in lung cancer prevention. *Cancer* **89**, 2506–2509 (2000).

8. Akhtar, N. & Bansal, J. G. Risk factors of lung cancer in nonsmoker. *Current problems in cancer* **41**, 328–339 (2017).

9. Biesalski, H. K. *et al.* European consensus statement on lung cancer: Risk factors and prevention. Lung cancer panel. *CA: a cancer journal for clinicians* **48**, 167–176 (1998).

10. Luo, G. *et al.* Projections of lung cancer incidence by 2035 in 40 countries worldwide: Population-based study. *JMIR public health and surveillance* **9**, e43651 (2023).

11. Wolf, A. *et al.* Data resource profile: Clinical practice research datalink (CPRD) aurum. *International journal of epidemiology* **48**, 1740–1740g (2019).

12. Head, A. *et al.* Inequalities in incident and prevalent multimorbidity in england, 2004–19: A population-based, descriptive study. *The Lancet Healthy Longevity* **2**, e489–e497 (2021).

13. Reeves, D. *et al.* Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: A cohort study with nested case–control analysis. *BMJ open* **4**, e004952 (2014).

14. Fry, J. S., Lee, P. N., Forey, B. A. & Coombs, K. J. How rapidly does the excess risk of lung cancer decline following quitting smoking? A quantitative review using the negative exponential model. *Regulatory Toxicology and Pharmacology* **67**, 13–26 (2013).

15. Samet, J. M. Lung cancer, smoking, and obesity: It's complicated. *JNCI: Journal of the National Cancer Institute* vol. 110 795–796 (2018).

16. Lubin, J. H., Caporaso, N., Wichmann, H. E., Schaffrath-Rosario, A. & Alavanja, M. C. Cigarette smoking and lung cancer: Modeling effect modification of total exposure and intensity. *Epidemiology* **18**, 639–648 (2007).

17. Reitsma, M. *et al.* Reexamining rates of decline in lung cancer risk after smoking cessation. A meta-analysis. *Annals of the American Thoracic Society* **17**, 1126–1132 (2020).

18. Navani, N. *et al.* Lung cancer in the united kingdom. *Journal of Thoracic Oncology* vol. 17 186–193 (2022).

19. O'Keeffe, L. M. *et al.* Smoking as a risk factor for lung cancer in women and men: A systematic review and meta-analysis. *BMJ open* **8**, e021611 (2018).

20. Jacob, L., Freyn, M., Kalder, M., Dinas, K. & Kostev, K. Impact of tobacco smoking on the risk of developing 25 different cancers in the UK: A retrospective study of 422,010 patients followed for up to 30 years. *Oncotarget* **9**, 17420 (2018).

21. Booth, H. P., Prevost, A. T. & Gulliford, M. C. Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for england, 2007 to 2011. *Pharmacoepidemiology and drug safety* **22**, 1357–1361 (2013).

22. Shiekh, S. I. *et al.* Completeness, agreement, and representativeness of ethnicity recording in the united kingdom's clinical practice research datalink (CPRD) and linked hospital episode statistics (HES). *Population Health Metrics* **21**, 3 (2023).

23. Spaulding, A. C. *et al.* Smoking in correctional settings worldwide: Prevalence, bans, and interventions. *Epidemiologic reviews* **40**, 82–95 (2018).

24. Richmond, R. *et al.* Tobacco in prisons: A focus group study. *Tobacco control* **18**, 176–182 (2009).

25. Dawkins, L. *et al.* A cross sectional survey of smoking characteristics and quitting behaviour from a sample of homeless adults in great britain. *Addictive behaviors* **95**, 35–40 (2019).

26. Garner, L. & Ratschen, E. Tobacco smoking, associated risk behaviours, and experience with quitting: A qualitative study with homeless smokers addicted to drugs and alcohol. *BMC Public Health* **13**, 1–8 (2013).

27. Fear, N. *et al.* Smoking among males in the UK armed forces: Changes over a seven year period. *Preventive medicine* **50**, 282–284 (2010).

28. Alberg, A. J. & Nonemaker, J. Who is at high risk for lung cancer? Population-level and individual-level perspectives. in *Seminars in respiratory and critical care medicine* vol. 29 223–232 (© Thieme Medical Publishers, 2008).

29. Liu, B. *et al.* Lung cancer in young adults aged 35 years or younger: A full-scale analysis and review. *Journal of Cancer* **10**, 3553 (2019).

30. Rich, A. *et al.* Non-small cell lung cancer in young adults: Presentation and survival in the english national lung cancer audit. *QJM: An International Journal of Medicine* **108**, 891–897 (2015).

31. Marston, L. *et al.* Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: A cross-sectional study. *BMJ open* **4**, e004958 (2014).

32. Polubriaginof, F., Salmasian, H., Albert, D. A. & Vawdrey, D. K. Challenges with collecting smoking status in electronic health records. in *AMIA annual symposium proceedings* vol. 2017 1392 (2018).