# MODERN EPIDEMIOLOGY

ASSIGNMENT 1

Hephaes Chuen Chau (M116 I54105029)

2022-02-25

## Question 0 (In-lesson exercise)

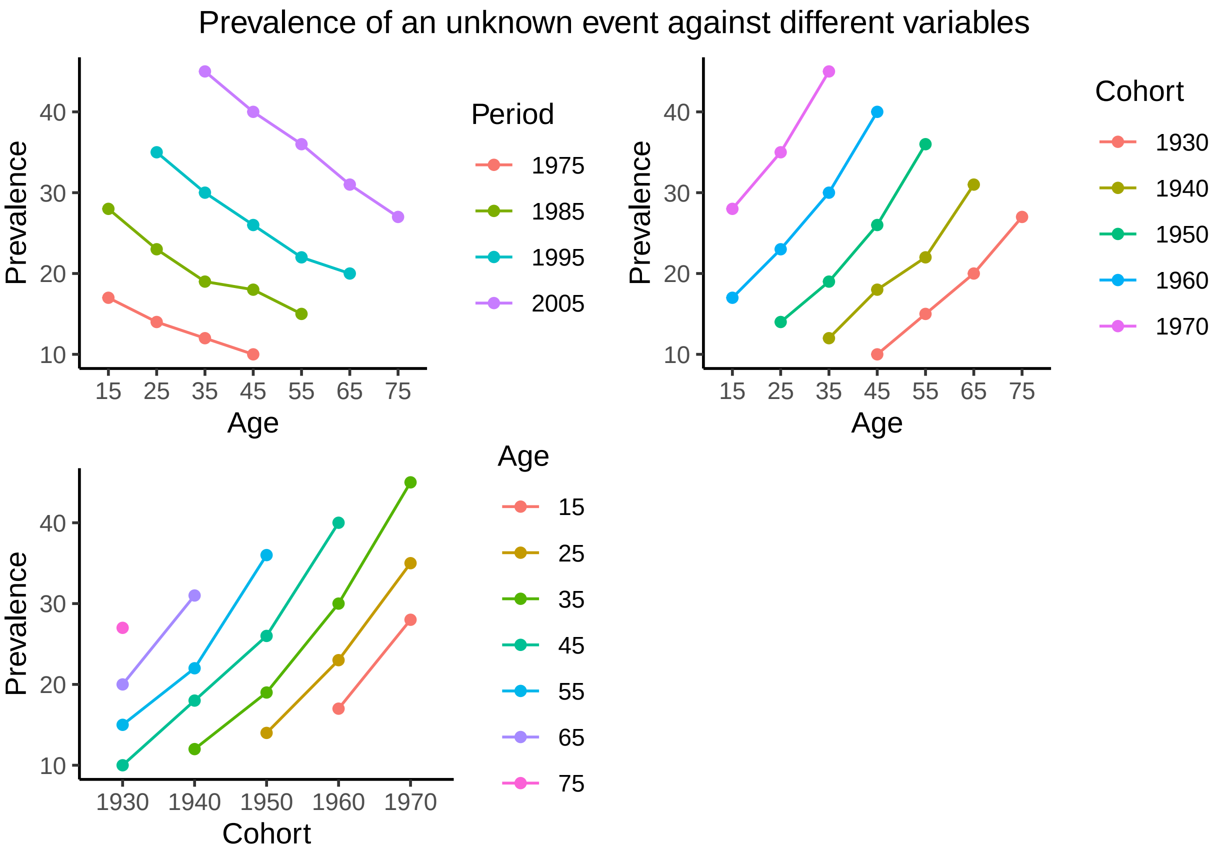
We begin by transforming the structure of the data:

```
## Rows: 7 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (5): Age, 1975, 1985, 1995, 2005
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 6 x 5
##     Age `1975` `1985` `1995` `2005`
##   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1    15     17     28     NA     NA
## 2    25     14     23     35     NA
## 3    35     12     19     30     45
## 4    45     10     18     26     40
## 5    55     NA     15     22     36
## 6    65     NA     NA     20     31
```

```
##   Cohort Age Period Prevalence
## 1   1960  15   1975         17
## 2   1970  15   1985         28
## 3   1950  25   1975         14
## 4   1960  25   1985         23
## 5   1970  25   1995         35
## 6   1940  35   1975         12
```

The plot is given as follows:



Explanations are given as follows:

1. We may first consider the prevalence vs age plot by period, if we fix a certain age group (eg, 45), it can be seen that the prevalence is higher in more recent period than in earlier period. This pattern repeats across age groups. Moreover, if we consider any period, the burden of the disease as it existed in that period is such that the the prevalence is higher in the younger populations.

2. We next consider the prevalence vs age plot by cohort. Here we see that for any given cohort, the prevalence of the disease increases with age. Therefore, it is not true that the likelihood of developing the disease decreases with age. What then explains the fact that more young people are observed to have the disease in any period? This can be explained from the plot again by the fact that for the younger cohorts, the starting prevalence at a young age is higher than in older cohorts, and thus since the increasing trend with age is not abolished for younger cohorts, in cross-sectional studies the young diseased members contribute to the increased prevalence observed at young age.

3. Finally, the prevalence vs cohort by age group reveals that for any given age group, the prevalence of the disease increases in younger cohorts. Summarising this observation with our previous observations, we can conclude that:

a. The likelihood of developing the disease increases with age, irrespective of birth year
b. Those who born later are more prone to be affected by the disease, irrespective of age
c. The burden of the disease is with the younger people, irrespective of period

## Question 1 (Exercise of Chapter 1)

### Question 1(a): After observing the incidence rates by age at any given year, it is concluded that "aging is not related to an increase in the incidence of Y and may even be related to a decrease in the incidence." Do you agree with this observation? Justify your answer.

I do not agree with this observation. If we identify the diagonal incidence rates in the table, we can see that the rates actually increase as calendar time is elapsed for a particular cohort. For example, for the cohort 1930-1934 (corresponding to those aged 20-24 in 1950), the rates increase as time progresses (10, 17, 20, …). **This suggests that a strong cohort effect is present despite the fact that** while a calendar time is fixed (eg, 1970), the rates are not distributed as such that higher age predicts a higher rate.

Alternatively, the same data as shown by the table is plotted with the $x$-axis being the birth cohort and the connected data points representing the data corresponding to a particular cohort. In such a plot, the specificity of the cohort effects is more apparent.

To do this we first transforms the structure of the data:

Before we have the data with the structure as:

```
## Rows: 8 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (9): Age, 1950, 1955, 1960, 1965, 1970, 1975, 1980, 1985
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

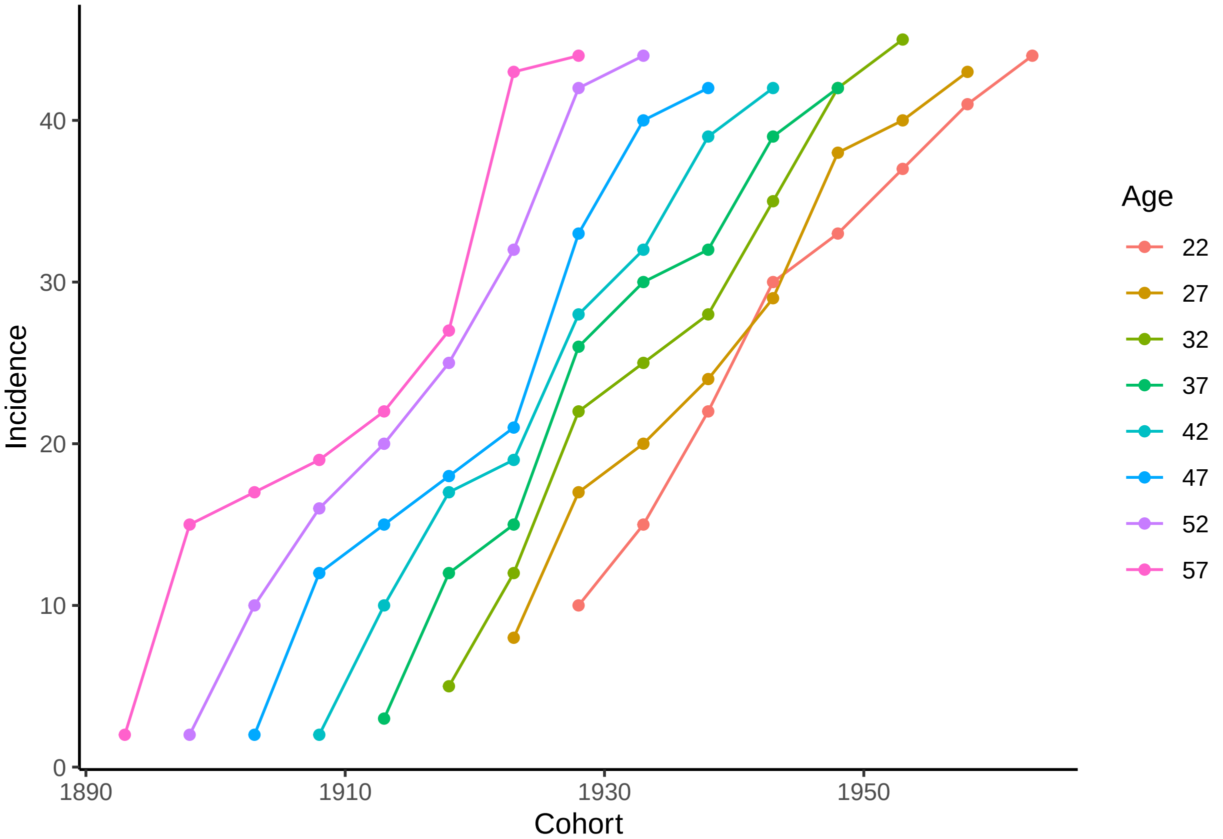
```
## # A tibble: 6 x 9
##     Age `1950` `1955` `1960` `1965` `1970` `1975` `1980` `1985`
##   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1    22     10     15     22     30     33     37     41     44
## 2    27      8     17     20     24     29     38     40     43
## 3    32      5     12     22     25     28     35     42     45
## 4    37      3     12     15     26     30     32     39     42
## 5    42      2     10     17     19     28     32     39     42
## 6    47      2     12     15     18     21     33     40     42
```

Now the data is transformed as

```
##   Cohort Age Period Incidence
## 1   1928  22   1950        10
## 2   1933  22   1955        15
## 3   1938  22   1960        22
## 4   1943  22   1965        30
## 5   1948  22   1970        33
## 6   1953  22   1975        37
```

We can thus plot the data (with each line corresponding to the same age group across different cohorts)

```
library(ggplot2)
library(ggrepel)
ggplot(new_df, aes(x = Cohort, y = Incidence,
                 group = Age)) +
    geom_line(aes(color = Age)) +
    geom_point(aes(color = Age)) +
    theme_classic()
```



Thus the plot very clearly shows that for most cohorts, incidence actually increases when individual in the cohort age. The cohort effect is shown by the increasing trend exhibited by most lines: for a fixed age group (eg, 57), incidence is higher in the younger cohorts than in older cohorts.

### Question 1(b): What are the purposes of examining birth cohort vis-à-vis cross-sectional rates?

Examining the incidence rates of a disease across birth cohorts help us uncover "cohort effects". **If cohort effects exist, the investigators are led to conjecture what types of life experiences that accumulated and that were characteristic of a particular birth cohort may be responsible for the observed change in incidence rates.** For a hypothetical example, if we identified that the birth cohort in 1930s have increased rates of depression relative to younger cohorts, the investigators may conjecture that war experiences may play a role in affecting the susceptibility of the individuals in the cohort to depression (or that the war experiences may have acted as selective pressure in favour of those prone to depression).

Secondly, **examining the cross-sectional rates reveals the differential burden of the disease for different age groups as the disease exists today.** This may be significant to formulating a correct and cost-effective intervention. Again as a hypothetical example, the knowledge that younger people (as opposed to the elderly) exhibit higher rates of depression may warrant policy interventions that are very likely of different types than what would be expected if it were the elderly that were more depressed.

## Question 2: A case-control study is conducted within a well-defined cohort. The reason for this is that expensive additional data collection is needed, and the budget is not sufficient to obtain these data from all cohort participants.

### Question 2(a): What type of case-control study within this cohort would be ideal to study multiple outcomes, and why is the alternative case-control design not recommended?

Case-cohort study is ideal for studying multiple outcomes, while nested case-control study is not. There are at least two reasons nested case-control study is not appropriate:

1. In nested case-control study, control is identified simultaneously whenever a case occurs, and the control is identified from the risk set, which is the set of all subjects in the cohort who are at-risk for the event of interest (meaning that they are not the case). **Thus, if our interest is not restricted to only one outcome, it is necessary to draw more than one control group from the cohort.** Since the control groups differ for different outcomes, we cannot infer odds ratio with respect to a factor between the group with two outcomes co-occuring and the control group, because there is not such a control. For case-cohort, since a subcohort is fixed and defined as the control regardless of the outcome being studied, it is possible to draw such an inference.
2. More importantly, as informed by Kim et al, nested-case control study is analysed most frequently using conditional logistic regression model, which is a statistical method that cannot handle the response variable being non-binary (Kim, 2014). Although the same study cited has developed a statistical method to allow for comparison of secondary outcomes in nested-case control study, such method is not the standard and require assumptions stated in the study.

Meanwhile, in case-cohort study, the control is selected as a subcohort (random sample of the total cohort) at baseline, and therefore we can reuse the same control for another outcome of interest.

### Question 2(b): In this cohort study, prevalent cases were not excluded at baseline, and thus, the investigators chose to use baseline data to examine associations between suspected risk factors andl preva1ent disease. What type of approach is this, and what are its main advantage and disadvantage?

The approach is case-cohort study. Advantages include:

1. As said in 2(a), in case-cohort study we can reuse the subcohort selected as control to serve as control for multiple cases. An example is provided in the book as ARIC cohort study by Dekker et al.
2. Since the control selected is a sub-cohort, and we have probed the risk factors of this subcohort, we can infer the distributions of risk factors for the cohort.

The disadvantages are:

1. Since the control was not immediately identified as the cases occured (ie, the control is not time-indexed), and they are identified before the cases, we cannot match the cases with subjects without the event at the time the cases occured, unlike in nested-case control study.
2. It is possible the exposure of the cases cannot be reliably ascertained because of recall bias, particularly when the exposure of interest include environmental variables.
3. When the exposure variable of interest potentially changes over time, since we do not re-examine the exposure of the control at a later time in case-cohort study, the estimate of the effects of the exposure could be biased.

# Question 3

## Question 3(a)

The most important concern is the possibility of over-matching resulting from smoking being a non-confounder. Whether over-matching in this case occurs depends on whether smoking causally increases the chance of exposure to air pollution.

1. Firstly, it is conceivable that smoking may very unlikely causes one's exposure to air pollution to increase or decrease, although it is conceivable that smoking increases (causally) the likelihood of developing the concerned respiratory cancer. Thus, this means smoking is unlikely to be a confounder. Since it is not a confounder, stratifying by smoking has no value in helping us uncover the possible relationship between air pollution and the outcome. This observation is less problematic when smoking is still statistically associated with air pollution (as a hypothetical case in point, when living in a polluted area predicts lower socioeconomic status which predicts a higher propensity to smoking), in which case the smoking is still a confounder.
2. Secondly, and more relevantly for this study, it is conceivable that the relationship between ethnic background and exposure to air pollution is so strong in the concerned population that after controlling for ethnic background, the variability of air pollution exposure becomes very low within strata. This means that the efficiency of the effect estimate is decreased (because the variance of the estimate will inevitably increase).
3. With respect to smoking again, it is conceivable that if smoking is a mediator in the causal pathway from air pollution to respiratory cancer  eg, residents in a highly polluted area may be more mindful of their health and therefore tend to quit smoking, while residents in less polluted area have a more perfunctory attitude), then matching for smoking introduces bias.

## Question 3(b)

It is not reasonable. There could be at least two arguments supporting this answer:

1. As advised by the textbook, little statistical power is gained beyond four or five controls per case
2. There are indeed situations in which statistical power gained beyond four or five controls can be quite substantial, as explained by Hennessy et al (Hennessy et al, 1999):

a. When the correlation of the case and control exposures is high

b. When the prevalence of exposure in the controls is low

But in either case, if we refer to the plot of statistical power vs control-per-case curves by correlation or prevalence (Fig 1 and 2 in Hennessy's study) the curves rapidly flatten off beyond 7. The gain of statistical power beyond 7 is minimal.

## Bibliography

Hennessy, S., W. B. Bilker, J. A. Berlin, and B. L. Strom. 1999. 'Factors Influencing the Optimal Control-to-Case Ratio in Matched Case-Control Studies'. American Journal of Epidemiology 149 (2): 195–97. https://doi.org/10.1093/oxfordjournals.aje.a009786.

Kim, Ryung S., and Robert C. Kaplan. 2014. 'Analysis of Secondary Outcomes in Nested Case-Control Study Designs'. Statistics in Medicine 33 (24): 4215–26. https://doi.org/10.1002/sim.6231.