**Practical Statistical analysis of a cohort study: answers**

Wednesday 23 November, 11.50am-12.40pm

The practical will use the paper of Crowe et al “Diet and risk of diverticular disease in Oxford cohort of European Prospective Investigation into Cancer and Nutrition (EPIC): prospective study of British vegetarians and non-vegetarians”. *British Medical Journal* 2011:Jul 19;343:d4131

Let’s assume that the representation below corresponds to the true exposure and outcomes during the follow-up of 10 participants to the Oxford cohort used for this article. These data have been created for the purpose of this practical and are not related to the results displayed in the paper.

|  |  |  |
| --- | --- | --- |
| Participant ID: |  | Follow-up time |
| 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| 1 |  |   |  |  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | H |   |
| 2 |  |   |  |   |   |   |   |  |   |   |   |   |   |   |   |   |   |   |   |   |   | OD |   |   |
| 3 |  |  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4 |  |   |  |   |   |   |   |   |   |  |   |   |   |   |   |   |   | D |   |   |   |   |   |   |
| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | H |   |   |   |   |   |   |  |
| 6 |  |   |  |   |   |  |   |   |   |   |   |   |   |   |   |   | H |   |   |   |   |   |   |   |
| 7 |  |  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 8 |  |   |  |   |   |   |   |   |   |   |   |  |   |   |   |   |   |   |   |   |   |   |   |   |
| 9 |  |   |  |  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | E |
| 10 |  |   |  |   |   |   |   |   |   |  |   |   |   |   |   |   |   |   | H |   |   |   |   |   |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *Legend* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| H |  |  Hospital diagnosis of diverticular disease |  |   |  Dietary questionnaire |  |  |
| D |  |  Death with mention of diverticular disease |   | Vegetarian (no meat or fish) |  |  |
| OD |  |  Death without mention of diverticular disease |   | Non-vegetarian |  |  |  |  |
| E |  |  Emigration |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Each horizontal line corresponds to one participant (participant ID from 1 to 10). The horizontal axis represents the number of years of follow-up starting in 1993 (enrolment questionnaire sent between 1993 and 1999 as stated page 2, 1st column of the paper) and ending in 2009 (December 2008 for SMR, March 2009 for HES and September 2009 for NHS central register as stated page 2, 2d column of the paper). Each vertical dotted line represents 1 year. Note that some participants changed their exposure status during the follow-up period. We will assume that all participants of this sub-sample were younger than 90 years at the end of their follow-up and that the participants enrolled before the 1st April 1997 were not English.

The questionnaires at baseline were send few months after the enrolment (this does not change the fact that the participants are part of the population at risk as soon as they are enrolled) and **all questionnaires assess the vegetarian status during the year preceding the questionnaire (meaning that the vegetarian status must be determined from the colour of the cell preceding the questionnaire).**

1/ What is the total number of person-year and the total number of cases in this sub-sample?

What is the total number of 100,000 persons-year and the total number of cases in the entire sample used in the paper?

How is it likely to influence the results obtained in the sub-sample?

Number of cases in the sub-sample=5

Number of cases in the entire sample=812

Number of person-year in the sub-sample = 13+10+16+5+9+6+16+11+14+6=106

Number of 100,000 person-year in the entire sample = 5.47 (547,312/100,000 as stated page 3, 1st column of the paper)

Given the very small number of cases and the total number of person-year, the analyses performed in the sub-sample will result in very imprecise estimate (wide confidence intervals). Therefore, results due to chance are much more likely to occur.

2/ Let’s first assume that the exposure status is known at inclusion only. Calculate the cumulative incidence ratio (CIR) of diverticular disease for those vegetarian at inclusion compared to those not vegetarian at inclusion.

A1=Number of cases vegetarian at inclusion=3

A0=Number of cases non-vegetarian at inclusion=2

N1=Number of vegetarian at inclusion=6

N0=Number of non-vegetarian at inclusion=4



3/ Interpret this estimate.

Participants who were vegetarian at baseline have the same risk of developing a diverticular disease compared to participants who were not vegetarian at baseline.

4/ Is this the most appropriate measure of relative risk in this context and why?

The calculation of a cumulative incidence ratio does not consider the duration of follow-up, which is a disadvantage. Besides, it also ignores the available information on change in exposure status. As information on follow-up (person-time) is available, calculation of the incidence rate ratio would be recommended under the current scenario.

5/ Still assuming that the exposure status is known at inclusion only, calculate the incidence rate ratio (IRR) for the occurrence of diverticular disease among participants vegetarian at inclusion compared to those non-vegetarian at inclusion.

We suggest filling the following table to facilitate the calculations for the following questions:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Participant ID: | Case (0/1) | Question number 5 | Question number 7 | Question number 9 |
| Using exposure at baseline only to determine V status | Using also exposure at FU to determine V status (V status=1 if V at both baseline and FU) | Using the time spend as a vegetarian (>80%) to determine the V status |
| V status(0/1) | PY  | V status(0/1) | PY  | % time spend veg | V status(0/1) | PY |
| V | NV | V | NV | V>80% | V≤80% |
| 1 |   |   |  |   |   |  |   |   |   |  |   |
| 2 |   |   |   |   |   |   |   |   |   |   |   |
| 3 |   |   |  |   |   |  |   |   |   |  |   |
| 4 |   |   |   |   |   |   |   |   |   |   |   |
| 5 |   |   |  |   |   |  |   |   |   |  |   |
| 6 |   |   |   |   |   |   |   |   |   |   |   |
| 7 |   |   |  |   |   |  |   |   |   |  |   |
| 8 |   |   |   |   |   |   |   |   |   |   |   |
| 9 |   |   |  |   |   |  |   |   |   |  |   |
| 10 |   |   |   |   |   |   |   |   |   |   |   |
| Total |   |   |   |   |   |   |   |   |   |   |   |
|   |   | IRR= |  |   | IRR= |  |   | IRR= |  |  |   |
| IRR: incidence rate ratio V: Vegetarian |
| PY: person-year NV: Non-vegetarian |

These are data from a prospective cohort study, and we have data on the follow-up time available. In this case, we can thus calculate the incidence of disease in the exposed and the unexposed separately and divide them to obtain a rate ratio:



We first need to count the total number of person-years of follow-up for the exposed and the non-exposed.

As we assumed that the exposure status is known at inclusion only, all the years of follow-up for participants changing their exposure status during follow-up will be attributed to the group they belonged at inclusion.

We thus get the following:

T1=Person-years among those exposed (i.e. vegetarian at inclusion)

=13+16+9+6+16+14=74

T0=Person-years among those not exposed (i.e. not vegetarian at inclusion)

=10+5+11+6=32

* Thus, the relative risk (in this case the incidence rate ratio) can be calculated as follows:



6/ Interpret this measure of relative risk.

Participants who were vegetarian at inclusion have a 35% lower rate of diverticular disease ( (1-0.65)x100 ) compared to participants who were not vegetarian at inclusion (i.e. the rate of diverticular disease in vegetarian at inclusion is 0.65 times the rate of those who are not vegetarian at inclusion)

7/ Let’s now take into account that the exposure status is also available after 5 years of follow-up (as stated page 3, 2d column of the paper) and calculate the new incidence rate ratio (IRR) for the occurrence of diverticular disease among participants vegetarian at both inclusion and follow-up compared to those non-vegetarian at inclusion or at follow-up.

The new number of cases in the exposed and non-exposed group is the following:

A1=2

A0=3

since the participant number 6 who did not report being vegetarian in the questionnaire sent after 5 years of follow-up is now classify as a case in the non-exposed group.

The new result for the number of person-years is the following:

T1=Person-years among those exposed (i.e. vegetarian at inclusion and follow-up)

=13+16+9+16+14=68

T0=Person-years among those not exposed (i.e. not vegetarian at inclusion or follow-up)

=10+5+6+11+6=38

* Thus, the incidence rate ratio can be calculated as follows:



8/ Interpret this measure of relative risk and the difference compared to the previous one. Is this result consistent with the result observed in the paper?

Participants who were vegetarian at both inclusion and after 5 years of follow-up have a 63% lower rate of diverticular disease compared to participants who were not vegetarian at inclusion or follow-up (i.e. the rate of diverticular disease in vegetarian at both inclusion and follow-up is 0.37 times the rate of those who are not vegetarian at inclusion or follow-up)

Taking into account the change of exposure during follow-up allow accounting for the stronger association between vegetarianism and the occurrence of diverticular disease in this subsample when vegetarianism is maintained during at least 5 years.

This result is not consistent with the result observed in the paper. As stated page 3, 2d column, “they was no evidence that the association between vegetarianism and diverticular disease differed according to duration of adherence to a vegetarian diet. In the fully adjusted model, the risk among vegetarian who had followed a vegetarian diet for more than 5 years was 0.72 (0.57 to 0.92) and for all other vegetarians the relative risk was 0.64 (0.42 to 0.97) compared with non-vegetarian”.

This could be due to the fact that

(i) our result in the sub-sample is due to chance given the very small sample size

(ii) the results displayed in the paper use the fully adjusted model and the results for the crude model (no adjustment) are not given.

The stronger negative association observed in the subsample for long-term vegetarian could be related to confusion factors.

The vegetarians following the diet during more than 5 years would be more likely to have characteristics that are negatively associated with the occurrence of diverticular disease (ie. being non-smoker, normal weight, having a higher education, living in a area non deprived, not having hypertension, hyperlipidemia, a long-term medical treatment, never used oral contraceptive or hormonal replacement therapy for women as displayed in table 3 of the paper) compared to vegetarian following the diet during less than 5 year.

9/ Let’s now assume that all the participants of the sub-sample kept the research team informed of any change of exposure during follow-up meaning that the complete picture of exposure is known for the 10 participants.

Assuming that a genuine vegetarian is a person who was vegetarian during at least 80% of the follow-up period, calculate the new incidence rate ratio (IRR) for the occurrence of diverticular disease among exposed (genuine vegetarians) compared to non-exposed (person being non-vegetarian or vegetarian during no more than 80% of the follow-up period).

The new number of cases in the exposed and non-exposed group is the following:

A1=1

A0=4

since the participant number 5 who followed a vegetarian diet during only 56% of his follow-up (5x100/9) is now classify as a case in the non-exposed group.

The new result for the number of person-years is the following:

T1=Person-years among those exposed (i.e. genuine vegetarian)

=13+16+16+14=59

T0=Person-years among those not exposed (i.e. person being non-vegetarian or vegetarian during no more than 20% of the follow-up period)

=10+5+9+6+11+6=47

* Thus, the incidence rate ratio can be calculated as follows:



10/ Interpret this measure of relative risk and the difference compared to the previous one.

Participants who followed a vegetarian diet during at least 80% of the follow period have a 80% lower rate of diverticular disease compared to participants who were not vegetarian or those who followed a vegetarian diet during no more than 20% of the follow-up period (i.e. the rate of diverticular disease in genuine vegetarian is 0.20 times the rate of those who are not genuine vegetarian)

Knowing the exact picture of exposure during follow-up allows strengthening the negative association between vegetarianism and the occurrence of diverticular disease even more in this subsample.

This example highlighted the effect that a miss-classification bias (because of lack of information or measurement error) can have on estimates: the less we used information on exposure, the less accurate our classification of exposed vs. non-exposed was and the more the estimates were attenuated (closer from null, i.e. no association).

11/ What is the limit of the type of estimates calculated so far?

Those estimates do not take into account the fact that the negative association observed between vegetarianism and the occurrence of diverticular disease can be due to other factors associated with both the exposure and the outcome (i.e. confusion variables).

12/ What type of model was used in the paper to estimate the association between vegetarianism and the occurrence of diverticular disease? What is the main assumption of this model and how can it be tested?

A Cox proportional hazards regression model was used to calculate hazard ratio (HR) as estimates of the relative risk for diverticular disease and 95% confidence intervals (CI), using age as the underlying time variable.

The main assumption is the proportional hazard assumption which states that the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate. This implies that the effect of one covariate on the outcome does not change over time.

Given the form of the hazard function for the Cox proportional hazard model:

**

where *t* is the time, *X* is the matrix of covariates and *β* is the vector of coefficients, the effect of time only enters in the baseline hazard functionnot in the covariate part of the model *βX*.

This can be tested looking at Kaplan Meier curves. If the predictor satisfies the proportional hazard assumption, then the curve of survival versus time for each value of the predictor should be parallel. This solution works best for time fixed covariates with few levels but does not work well for continuous predictor or categorical predictors that have many levels because the graph becomes to "cluttered".

Another solution is therefore to create interactions between the predictors and a function of survival time and include them in the model. If any of the time dependent covariates are significant then those predictors are not proportional.

13/ What does mean “All analyses are stratified by sex, method of recruitment, and region of residence)”?

That means that a different baseline hazard functionis created for each combination of sex, method of recruitment, and region of residence. Consequently, there are 8 different baseline hazard functions in the present study

8=2 (categories for sex: Male or Female) x 2 (categories of methods of recruitment: General practice or Post) x 2 (categories for region of residence: England or Scotland)

Performing stratified analyses (ie. estimating different baseline hazard functions) permits to reduce the chance of proportional hazard failure.

14/ What were the adjustment variables?

The adjustment variables were smoking, education level, Townsend deprivation index, self reported hyperlipidaemia, receipt of long term treatment for any illness, use of oral contraceptive, use of hormone replacement therapy and BMI.

15/ Interpret the Hazard ratio of 0.70 (0.56 to 0.87) displayed in the table 4 (1st row, 2d column).

Participants who were vegetarian at inclusion (ie not eating meat or fish) had a 30% (13% - 44%) reduced risk of diverticular disease compared to participants not vegetarian at inclusion. This is equivalent to say that the hazard rate of diverticular disease in vegetarian at inclusion is 0.70 times the hazard rate in non-vegetarian at inclusion. This association was independent of smoking, education level, Townsend deprivation index, self reported hyperlipidaemia, receipt of long term treatment for any illness, use of oral contraceptive, use of hormone replacement therapy and BMI.

16/ Write the detailed formula of the model used to test the interaction between diet group and sex (table 6).

**

where **are the 8 baseline hazard functions

*Vegetarian*=1 if the participant is vegetarian at inclusion and 0 otherwise

*Formersmoke*=1 if the participant is a former smoker at inclusion and 0 otherwise

*Lightsmoke*=1 if the participant is a light smoker at inclusion and 0 otherwise

*Heavysmoke*=1 if the participant is a heavy smoker at inclusion and 0 otherwise

*Higher2deducation*=1 if the highest degree of the participant is secondary school and 0 otherwise

*University*=1 if the highest degree of the participant is university and 0 otherwise

*Cat2deprivationindex*=1 if the participant live in a town with a deprivation index between -3.3 and -1.7 and 0 otherwise

*Cat3deprivationindex*=1 if the participant live in a town with a deprivation index between -1.8 and 0.2 and 0 otherwise

*Cat4deprivationindex*=1 if the participant live in a town with a deprivation index of at least 0.3 and 0 otherwise

*Hyperlipidemia*=1 if the participant reported a hyperlipidelia at inclusion and 0 otherwise

*Longtermtreatment*=1 if the participant reported a long term treatment at inclusion and 0 otherwise

*Oralcontraceptive*=1 if the participant (women) reported ever using oral contraceptive and 0 otherwise

*HRT*=1 if the participant (women) reported ever using HRT and 0 otherwise

*BMI* is the continuous value of BMI for each participant.

*Sex*=1 if the participant is vegetarian a men and 0 otherwise

Note that the reference groups (for example never smoker in the case of the smoking status) are not entered in the model since the information can be deduced from the other dichotomous variables (never smokers are participants with 0 value for *Formersmoke*, *Lightsmoke* and *Heavysmoke*)

17/ What is the difference of interpretation between the HR of BMI and the HR of light smoker?

BMI is a continuous variable (the only one of the model) whereas light smoker is a dichotomous variable which equals 1 for participant reporting being light smoker at inclusion and 0 otherwise.

This means that the HR for the BMI variable can be interpreted as the factor by which the hazard rate of diverticular disease will be multiplied for each unit of BMI increase.

The HR of the light smoker dichotomous variable can be interpreted as the factor by which the hazard rate of diverticular disease will be multiplied when a participant is a light smoker compared to never smoker.

18/ What is the interpretation of the estimate obtained for the interaction term diet group x sex (table 6)?

Adding an interaction term in the model permits to obtain HR for a combination of variables instead of HR for each variable separately

Let’s have a look at the interpretation of the three coefficients (β1, β15, β16) depending of the value of the two variables *Vegetarian* and *Sex:*

1/ If the participant is female and non-vegetarian: *Vegetarian* and *Sex* both equal 0 which leads to: β1 x 0 + β15 x 0 + β16 x 0 x 0 = 0.

This is the reference group.

2/ If the participant is female and vegetarian: *Vegetarian*=1 and *Sex*=0 which leads to:

β1 x 1 + β15 x 0 + β16 x 1 x 0 = β1 .

This means β1 corresponds to the effect of being vegetarian in women compared to not being vegetarian (reference group).

3/ If the participant is male and non-vegetarian: *Vegetarian*=0 and *Sex*=1 which leads to: β1 x 0 + β15 x 1 + β16 x 0 x 1 = β15 .

This means β15 corresponds to the effect of being male in non-vegetarian compared to being female in non-vegetarian (reference group).

4/ If the participant is male and vegetarian: *Vegetarian*=1 and *Sex*=1 which leads to:

β1 x 1 + β15 x 1 + β16 x 1 x 1 = β1 + β15 + β16.

We already know that β1 corresponds to the effect of being vegetarian in women and β15 corresponds to the effect of being male in non-vegetarian. This implies that β16 is the difference of the effect between being vegetarian in male compared to being vegetarian in women. Summing β1 andβ16 permits to obtain the effect of being vegetarian in men.

Therefore, if β16 is significantly different from 0 (we look at the coefficient here, not the HR), this means that the effect of being vegetarian in male is significantly different from the effect of being vegetarian in women.

In the present paper, it is not the case since P=0.699 for interaction which means that 0.65 (0.41 to 1.02) is not significantly different from 0.71 (0.55 to 0.92).