## Simulated Outbreak Exercise

|  |  |
| --- | --- |
| Scenario Date: | Tuesday 10 October (16:00-17:00) |
| Inject No: | 10 |
| Inject time: | 1 hour |
| Inject Title: | Line list Case definition |

1. **Learning outcomes:**At the end of the session, participants will be able to:

- Apply case definition criteria to a dataset using logical arguments in R

1. **Story/plot description:**  
   You will now create a new column in the data set to hold the case definition you decided in a previous step during your investigation. You can call this column case and set it to TRUE if the individual meets the case definition criteria and FALSE if not. You will use this column later on for any calculations needed (descriptive statistics, two-by-two tables to compute measures of association, etc.) to figure out the culprit of this outbreak.
2. **Questions/assignments for the group:**

If you closed your R project, please open it again and load the clean dataset. We strongly recommend you use the “Copenhagen\_clean1.rds” provided by us in EVA. This is just to be sure all needed changes in that file for the code to work have been done (as you may have changed the file in a slightly different way than we have).

To be sure we all start from the same case definition, let’s agree that a case was defined as a person who:

* attended the school dinner on 11 November 2006 (i.e. is on the linelist)
* ate a meal at the school dinner (i.e. was exposed)
* fell ill after the start of the meal
* fell ill within the time period of interest after the school dinner
* suffered from diarrhoea with or without blood, or vomiting

Non cases (not-ill) were defined as people who:

* attended the school dinner on 11 November 2006 (i.e. are on the linelist)
* ate a meal at the school dinner (i.e. were exposed)
* did not fall ill within the time period of interest
* did not develop diarrhoea (with or without blood) or vomiting

3.1. Install packages (if needed) and load libraries.

3.2. Import your data.

3.3. Identify the variables you need to apply the case definition criteria.

Facilitator’s note: The variables we need from the dataset to apply the above case definition are: meal, onset\_datetime, diarrhoea, bloody and vomiting.

The meal is our exposure of interest at this point. Note that not all participants who attended the party actually ate their meal! See summary(copdata$meal)

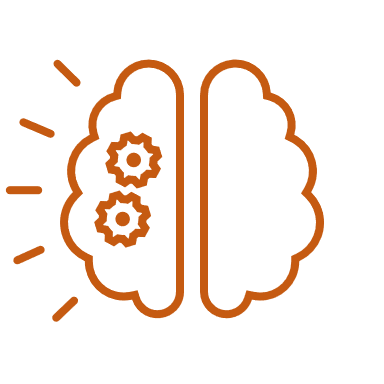
3.4 Create a new case column to hold the binary case definition variable. Let’s think about how to do this little by little:

a) Ate a meal at the school dinner (i.e. was exposed).

* You decide to exclude any people from the cohort who didn’t eat at the dinner, because we specifically hypothesised a food item to be the vehicle of infection in this outbreak. Thus, filter your dataset to those who ate a meal.



**Hint**: filter by those with meal == TRUE.



**Stop and reflect**: What are some of the implications this decision may lead to?

(excluding any people from the cohort who didn’t eat at the dinner)

Facilitator’s note: Seven of the respondents actually said they did not eat the meal, but when it came to the questions about which food items they ate, they provided answers! This issue could have been minimised by:

* At the survey state, one could adjust the design of an electronic questionnaire to prevent key questions from being skipped. This can come with both pros and cons. Allow fellows to discuss if time allows.
* Explore your data further, realise this is the case, and recode the meal variable for these individuals as TRUE. 🡺 This would be the way to go, but is not what we did in our example because we tried to keep it simple, and also because it is good to show that you may not always clean the data perfectly, and that has consequences: You can highlight the importance or really explore your data in depth.

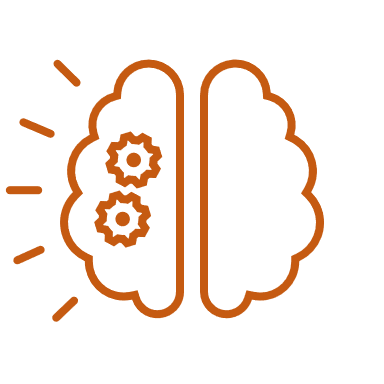
By making the above decision, we may be missing cases and non-cases people, and thus, modifying the final estimate of our measure of association. 🡺 It is very important to know your data, explore it deeply and try to clean it as well as possible. Every step one makes when cleaning the data may have a consequence, and we should be aware of it when making the data cleaning decisions and when interpreting the results.

b) Fell ill after the start of the meal.

* We define “fell ill” as any person who reported having diarrhoea with OR without blood, OR vomiting. To capture this information easily, you will create a new gastrosymptoms variable. This variable will indicate that the person had one OR (“or” is R is achieved by using |) more of the clinical symptoms in your definition.



**Hint**: Create a new gastrosymptoms column in your line list with mutate() and case\_when().



**Stop and reflect**: What are some of the implications this decision may lead to?

(defining “fell ill” as any person who reported having diarrhoea with OR without blood, OR vomiting)

Facilitator’s note: Having one clinical symptom enough to be considered a potential case at this point may be considered too unspecific (low specificity). For example, a person who ate at the dinner party and developed diarrhoea for other reasons other than food poisoning (say, they recently started on antibiotics known for unbalancing the intestinal flora and causing diarrhoea) could be misclassified as a potential case. (Note we talk about ***potential* case**, and not **case;** that is because here we are not talking about cases per-se yet, but this decision has implications for when applying the case definition below).

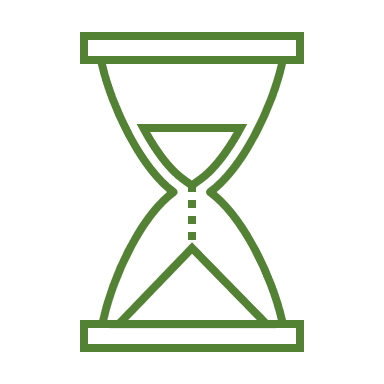
Moreover, those who did not report clinical symptoms will be defined as non-cases. Thus, we are assuming that these individuals did not develop symptoms because they didn’t report them. The missing values could be due to, for example, them skipping the questions in the questionnaire. Some individuals may be reluctant to report symptoms, due to shame, fear of repercussion, or others. It is important to think ahead, before the interview, about ways to minimise these situations. For example, through questionnaire design, you may impede skipping questions; you could promote trust by using the right interviewers (in some cases this will be someone from the community, in others someone form specific NGOs, someone of a specific race or gender, etc); choose to carry out online questionnaires vs in person (or *vice versa*, depending on the situation), etc.

c) Fell ill within the time period of interest.

* + - Hmmm… what is the time period of interest? You could start calculating the incubation period, which can be defined by calculating the time between exposure (the meal) and onset of symptoms, and then looking at the distribution of these time differences. In this outbreak, incubation periods are easy to calculate, because everyone was exposed at (roughly) the same time and on the same day (eating the meal at the school dinner party).
    - Dinner was served at 18:00 roughly for everyone. Create a new meal\_datetime variable with this information for all people in your dataset.
    - You can calculate the incubation period for each participant by subtracting onset\_datetime - meal\_datetime.
    - Then, take the median() of that column to calculate a median incubation period, this will help you start having some hypothesis about the type of pathogen we are dealing with until the lab results come back.

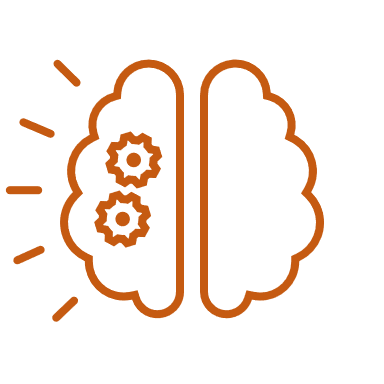


**Hint**: Be aware of the type of variable incubation is (check class()), as well as of missing values NA.

*Note: The* ***incubation period*** *of a disease is the time from an exposure resulting in infection until the onset of disease. Because infection usually occurs immediately after exposure, the incubation period is generally the duration of the period from the onset of infection to the onset of disease – Rothman, Greenland, Lash (2017): Modern Epidemiology, 3rd edition*

Facilitator’s note: If you pay attention, there were two individuals who don’t have a time recorded, only a date (Nov 11), the day of the dinner. These people probably got sick the same night they had the dinner, but the exact time was not recoded. They way we’ve managed the data will insert “00:00” in a missing value of a time. This means that the 2 people that got sick the day of the dinner have been recorded sick *before* the dinner (in the early morning of Nov 11). The two people have a negative incubation period! This should not happen. Advanced fellows can modify this and fix the error (if they do, their results will be a bit different, but similar). Beginners can continue with this “erroneous” data, for easiness.

* + - Based on results, say you define and limit the maximum incubation period to 48 hours after the meal, as the data points to a fast-acting bacterial toxin or a virus. That is, they should have developed (onset\_datetime) at least one symptom (diarrhoea, bloody or vomiting) 48h after meal\_datetime to become a case.

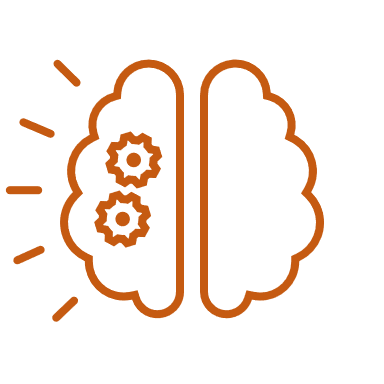


**Stop and reflect**: What are some of the implications this decision may lead to? (implications of this case definition)

Facilitator’s note: All those not developing at least one symptom (diarrhoea with OR without blood, OR vomiting) 48h after the dinner are considered non-cases. This could (depending on how you decide to analyse your data) include those who had no symptoms at all, those who have missing data on the onset\_datetime variable, and/or those who had symptoms before eating the meal\*. This is a reminder that you need to be both careful and aware of the implications of your data analysis decisions.

\*If a person had clinical symptoms before eating the meal, they are considered as not-cases. However, it could be that a person had symptoms before the meal, and yet, still got infected by the pathogen when eating their meal (bad luck, we know…). According to this definition, we would be missing that case.

d) Finally, with this information you can create a new case column to hold the binary (TRUE/FALSE) case definition variable.



**Stop and reflect**: What do you think are the risks of mis-classifying cases as non-cases in your analysis?

Facilitator’s note: we will have bias either towards or away from the null, depending on the proportions of subjects misclassified.

3.5 Export clean data.

* Save your clean data as a new file “Copenhagen\_clean2\_YOURINITIALS”, under the data folder.

1. **Notes for facilitators:**

If the fellow’s laptop’s language settings are not English, it might not recognize the `ymd\_hm` input. They can either use a `ymd\_h` or `ymd\_hms` input, or switch their settings to English by running:

# Check current system settings:

Sys.getlocale()

# Set time “in English”:

Sys.setlocale("LC\_TIME", "English")

Notes on using *case\_when()*:

* Beware of missing values; when using case\_when() these need to be explicitly categorised, otherwise they will be missed out. We know that the data collection method was an online questionnaire, so it is possible that that some people skipped the questions on symptoms, when they didn’t have any. If a respondent had no gastrointestinal symptoms, but did attend the meal, we would want to include this person in the study as a non-case.
* When constructing the case\_when() statements, we start with the most specific condition (where all the symptoms of interest are not equal to TRUE) and ended with the most general condition (where any of the symptoms of interest are present). Note that if you build your case\_when() conditions in the opposite order (most general to most specific), you will get different results. Remember that if two consecutive logical conditions overlap, the last one will overwrite the preceding one.

1. **Scenario for the role play:**  
   N/A
2. **Timing:**  
   60 minutes
3. **Need for materials (logistics)**
4. **Deliverables**  
   A script with reproducible code and clean dataset (Copenhagen\_clean2\_YOURINITIALS.rds) with added case definition column named case.