Practical I: epidemiological surveillance and modelling (non-spatial)

29th/30th January 2026

In this first practical the student is expected to learn to download, visualise, analyse, and interpret epidemiological case count data. Further, the student is expected to reflect on the issues these data present and provide a perspective on how to improve disease surveillance and analytics in order to better estimate basic epidemiological parameters.

The practical will start with a short 15-20 minute presentation by the course convenor and demonstrators. The next 2h will be spent answering predefined questions before a short wrap up and reflection at the end.

There will be a 15 minute break after ca. 1h ½.

```r
library(knitr)
library(rmarkdown)
library(dplyr)
library(EpiEstim)
library(ggplot2)
library(incidence)
```

Practical I Questions

Q1: What are common disease surveillance data and what are their pros and cons? Name 4 at most. (ca. 15 minutes)

Q2: What are some of the challenges with public health surveillance at the beginning of disease outbreaks? (ca. 5 minutes)

Q3: Describe, in your own words, what the time-varying reproduction number Rt represents? Further, list the data sources necessary to estimate Rt? Focus on main manuscript and high level insights (ca. 20 minutes). Reading: Anne Cori, Neil M. Ferguson, Christophe Fraser, Simon Cauchemez, A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics, American Journal of Epidemiology, Volume 178, Issue 9, 1 November 2013, Pages 1505–1512, https://doi.org/10.1093/aje/kwt133 (https://doi.org/10.1093/aje/kwt133)

Q4: Downloading national COVID-19 cases data (England).

- Historical data related to COVID-19 in the UK can be downloaded from the UKHSA data dashboard https://ukhsa-dashboard.data.gov.uk/covid-19-archive-data-download (https://ukhsa-dashboard.data.gov.uk/covid-19-archive-data-download)

- Given the size of the data, a pre-processed version of the relevant case data has been prepared for you and can be read into R directly using the link: https://raw.githubusercontent.com/kraemer-lab/Mbiol-Practical-2026/refs/heads/main/Data/nation_newCasesBySpecimenDate.csv (https://raw.githubusercontent.com/kraemer-lab/Mbiol-Practical-2026/refs/heads/main/Data/nation_newCasesBySpecimenDate.csv)

- Make sure that you understand the structure of the data and the variables it contains.

- For this practical we are only interested in cases in England between 1st October 2021 and 31st December 2021. So restrict your dates and sort them from October to December 2021:

```
dat <- dat %>%
  mutate(date = as.Date(date)) %>%
  arrange(date) %>%
  filter(date >= "2021-10-01", date <= "2021-12-31", area_name == 'England') %>%
  mutate(t_end = 1:n())
```

Q5: Visualising COVID-19 cases data.

- Use ggplot and plot cases by date.

```
ggplot(data = dat, aes(x = date, y = value)) +
  geom_point() +
  scale_x_date(date_breaks = "2 week") +
  theme_bw()+
  labs(y = "New Number of Cases (Specimen Date)")
```

NOTE - In reality we would remove the weekend effect.

Q6: Estimating time-varying reproduction number, Rt.

We can run `estimate_R` on the incidence data to estimate the time-varying reproduction number Rt. For this, we need to specify i) the time window(s) over which to estimate Rt and ii) information on the distribution of the serial interval (SI).

For i), the default behavior is to estimate Rt over weekly sliding windows (for example: window 1 = [01Oct2021, 07Oct2021], window 2 = [02Oct2021, 08Oct2021], window 3 = [03Oct2021, 09Oct2021], etc). This can be changed through the `t_start` and `t_end` arguments (see below, “Changing the time windows for estimation”). For ii), there are several options, specified in the method argument.

The simplest is the `parametric_si` method, where we only specify the mean and standard deviation of the SI.

In this example, we only specify the mean and standard deviation of the serial interval. In the following example, we use the mean (4.1 days) and standard deviation (2.8) of the serial interval for SARS-CoV-2 Delta variant of concern from Backer et al. 2022: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2022.27.6.2200042 (https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2022.27.6.2200042)

Make sure the date are in chronological order, because `estimate_R` requires daily cases from earliest date.

After estimating Rt, plot the output. (ca. 10 minutes)

```
res_parametric_si <- estimate_R(dat$value, method="parametric_si", config = make_conf
ig(list(mean_si = 4.1, std_si = 2.8)))

# Extract the estimates time-varying reproduction numbers
res_R <- res_parametric_si[["R"]]

# Add dates from the original dataset
res_R <- merge(res_R, dat, by = c("t_end"))

# Plot median and 95% uncertainty intervals
ggplot(data = res_R, aes(x = date, y = `Median(R)`)) +
  geom_line() +
  geom_ribbon(aes(ymin = `Quantile.0.025(R)`, ymax = `Quantile.0.975(R)`), alpha = 0.
4) +
  geom_hline(yintercept = 1, color = "blue") +
  scale_x_date(date_breaks = "2 week") +
  theme_bw()
```

Q7: Based on the above plot, describe the epidemiological dynamics of the outbreak by referencing the change in the time-varying reproduction number, Rt. (ca. 10 minutes)

Q8: Estimating variant-specific case incidence.

Between October and December 2021, the previously dominant Delta variant was rapidly displaced by the Omicron variant. It is therefore important that we estimate the time-varying reproduction number for each variant separately. For this, we need to first estimate the case incidence attributed to each variant using lineage proportion estimates from genomic surveillance data. In this example, we will use publically available data from the Sanger Institute which can be inspected here (https://covid19.sanger.ac.uk/lineages/raw (https://covid19.sanger.ac.uk/lineages/raw)). A pre-processed version of this data has already been prepared for you and can be read into R directly using the below link: https://raw.githubusercontent.com/kraemer-lab/Mbiol-Practical-2026/refs/heads/main/Data/daily_lineage_freqs.csv (https://raw.githubusercontent.com/kraemer-lab/Mbiol-Practical-2026/refs/heads/main/Data/daily_lineage_freqs.csv), following a similar procedure as in Q4.

Visualise the data and describe in your own words (ca. 10 minutes).

```
lineage_freqs.df <- read.csv('https://raw.githubusercontent.com/kraemer-lab/Mbiol-Pra
ctical-2026/refs/heads/main/Data/daily_lineage_freqs.csv') %>%
  mutate(date=as.Date(date))
ggplot() +
  geom_line(dat=lineage_freqs.df, aes(x=date, y=proportion, color=VOC)) +
  theme_bw()
```

To estimate the incidence of each variant separately, we have to multiple the daily case incidence (from Q1-2) by the proportion of cases attributed to each variant from the lineage frequency file. Visualise the variant-specific incidence and describe what you see (ca. 10 minutes).

```
omicron.incidence.df <- merge(dat, lineage_freqs.df[lineage_freqs.df$VOC == 'Omicro
n',], by=c('date')) %>%
  mutate(estCases=value*proportion)

delta.incidence.df <- merge(dat, lineage_freqs.df[lineage_freqs.df$VOC == 'Delta',],
by=c('date')) %>%
  mutate(estCases=value*proportion)

combined.incidence.df <- rbind(
  omicron.incidence.df %>% select(date, estCases) %>% mutate(VOC='Omicron'),
  delta.incidence.df %>% select(date, estCases) %>% mutate(VOC='Delta')
)

ggplot() +
  geom_point(dat=combined.incidence.df, aes(x=date, y=estCases, color=VOC)) +
  scale_x_date(date_breaks="2 week") +
  labs(x='date', y='estimated daily case incidence') +
  theme_bw()
```

Do you anticipate any potential issues if we are to estimate Rt for Omicron for the same time period as in Q6? (ca. 5 minutes)

Q9: Estimate variant-specific time-varying reproduction number, Rt.

Before we go ahead and run `estimate_R` on the variant-specific incidence data, we have to first make sure that the parameters we use to describe the distribution of the serial interval (SI) is appropriate for each variant. There are several studies estimating SI for both the Delta and Omicron variants (e.g. Backer et al., Eurosurveillance, 2022 (https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2022.27.6.2200042); Águila-Mejía et al., Emerging Infectious Diseases, 2022 (https://pmc.ncbi.nlm.nih.gov/articles/PMC9155885/)). In this example, we will use a mean serial interval of 4.1 days (SD: 2.8 days) and 3.5 days (SD: 2.4 days) for the Delta and Omicron variants, respectively.

What is the significance of a shorter serial interval? (ca. 5 minutes)

Now having both the variant-specific incidence data and the SI parameters, we can estimate Rt for each variant separately by following the same steps as in Q6. Uncomment (remove `#` symbol) and enter an appropriate earliest cut-off date of the time window as determined in Q8. Make sure to replace the first argument in `estimate_R` with the correct dataframe and column name. Visualise the results and describe what you see. (ca. 10 minutes)

```r
# left_limit.date <- as.Date("<cut-off-date>")

omicron_res_parametric_si <- estimate_R(omicron.incidence.df[omicron.incidence.df$date >= left_limit.date,]$estCases, method="parametric_si", config = make_config(list(mean_si = 3.5, std_si = 2.4)))
# Extract the estimates time-varying reproduction numbers
omicron_res_R <- omicron_res_parametric_si[["R"]]
# Adjust t_end
omicron_res_R <- omicron_res_R %>% mutate(t_end = t_end + as.integer(left_limit.date - min(dat$date)))
# Add dates from the original dataset
omicron_res_R <- merge(omicron_res_R, dat, by = c("t_end"))

delta_res_parametric_si <- estimate_R(delta.incidence.df[delta.incidence.df$date >= left_limit.date,]$estCases, method="parametric_si", config = make_config(list(mean_si = 4.1, std_si = 2.8)))
# Extract the estimates time-varying reproduction numbers
delta_res_R <- delta_res_parametric_si[["R"]]
# Adjust t_end
delta_res_R <- delta_res_R %>% mutate(t_end = t_end + as.integer(left_limit.date - min(dat$date)))
# Add dates from the original dataset
delta_res_R <- merge(delta_res_R, dat, by = c("t_end"))

# Combined dataframes
combined_res_R <- rbind(omicron_res_R %>% mutate(VOC='Omicron'), delta_res_R %>% mutate(VOC='Delta'))

# Plot median and 95% uncertainty intervals
ggplot(data = combined_res_R, aes(x = date, y = `Median(R)`, color = VOC)) +
  geom_line() +
  geom_ribbon(aes(ymin = `Quantile.0.025(R)`, ymax = `Quantile.0.975(R)`), alpha = 0.4) +
  geom_hline(yintercept = 1, color = "blue") +
  scale_x_date(date_breaks = "2 week", limits=c(left_limit.date, as.Date('2021-12-31'))) +
  scale_y_continuous(limits = c(0, 10)) +
  theme_bw()
```

Describe the patterns in the Rt estimates for each variant. Can you explain them? (ca. 10 minutes)

Q10: Accounting for case importation (Omicron variant).

We have so far assumed that all cases are linked by local transmission. However, it was later discovered through analysis of travel history data that a substantial proportion of cases of the Omicron variant were in fact imported from abroad [see examples of analyses that make use of such data here: du Plessis et al. (https://www.science.org/doi/10.1126/science.abf2946) and Tsui et al. (https://www.science.org/doi/10.1126/science.adg6605)]. A dataset containing the proportion of Omicron cases that were imported on each day is available for you and can be read into R directly by copying and pasting the link: https://raw.githubusercontent.com/kraemer-lab/Mbiol-Practical-2026/refs/heads/main/Data/daily_import_proportions.csv (https://raw.githubusercontent.com/kraemer-lab/Mbiol-Practical-2026/refs/heads/main/Data/daily_import_proportions.csv) as you did in Q8. We can use this information to account for case importation in our Rt estimates.

Use the data to estimate the number of Omicron cases that were local/imported on each day. Visualise the results and describe what you see. Can you explain the patterns? (ca. 15 minutes)

```
import.proportion.df <- read.csv('https://raw.githubusercontent.com/kraemer-lab/Mbiol-Practical-2026/refs/heads/main/Data/daily_import_proportions.csv') %>%
  mutate(date=as.Date(date))

adjusted.incidence.df <- omicron.incidence.df %>%
  select(date, estCases) %>%
  mutate(importProportion = import.proportion.df$prop) %>%
  mutate(imported = as.integer(estCases*importProportion)) %>%
  mutate(local = estCases - imported)

ggplot() +
  geom_point(dat=adjusted.incidence.df, aes(x=date, y=local), color='black') +
  geom_point(dat=adjusted.incidence.df, aes(x=date, y=imported), color='red') +
  scale_x_date(date_breaks="2 week") +
  labs(x='date', y='estimated daily number of local/imported cases') +
  theme_bw()
```

Having adjusted for case importation, we can now re-estimate Rt for the Omicron variant following the same steps as before. Visualise both the original and adjusted Rt estimates and describe what you see. Is there a difference following the adjustment, and if yes, can you explain why? (ca. 10 minutes)

```
omicron_res_with_imports <- estimate_R(adjusted.incidence.df[adjusted.incidence.df$date >= left_limit.date,], method = "parametric_si", config = make_config(list(mean_si = 3.5, std_si = 2.4)))
omicron_res_with_imports_R <- omicron_res_with_imports[["R"]]
omicron_res_with_imports_R <- omicron_res_with_imports_R %>% mutate(t_end = t_end + as.integer(left_limit.date - min(dat$date)))
omicron_res_with_imports_R <- merge(omicron_res_with_imports_R, dat, by = c("t_end"))

ggplot() +
  geom_line(dat = omicron_res_R, aes(x = date, y = `Median(R)`), color='black') +
  geom_line(dat = omicron_res_with_imports_R, aes(x = date, y = `Median(R)`), color='red') +
  geom_ribbon(dat = omicron_res_R, aes(x = date, ymin = `Quantile.0.025(R)`, ymax = `Quantile.0.975(R)`), alpha = 0.4, fill='black', color='black') +
  geom_ribbon(dat = omicron_res_with_imports_R, aes(x = date, ymin = `Quantile.0.025(R)`, ymax = `Quantile.0.975(R)`), alpha = 0.4, fill='red', color='red') +
  geom_hline(yintercept = 1, color = "blue") +
  scale_x_date(date_breaks = "2 week", limits=c(left_limit.date, as.Date('2021-12-31'))) +
  scale_y_continuous(limits=c(0, 9)) +
  theme_bw()
```

Q11: Name other sources of uncertainty in Rt estimation?