Statistical modelling in public health: SARS-CoV-2 – Analysis of Epidemiological Parameters

Dr. School seminar, Summer 2024

Lukas Richter

17.05.2024

Background

- First SARS-CoV-2 case reported end of February 2020
- Early pandemic (outbreak)
  - Urgent need to inform
  - Severity? (hospitalisation, case-fatality rate, mortality)
  - Speed of spread? (reproduction number, serial interval)
- Mid pandemic
  - Data imputation
  - changing variants
- Late pandemic
  - (unfamiliar) huge amounts of data/cases

Serial interval

Definition *SI*

The serial interval (SI) is the **time between disease onset** of successive cases in a chain of transmission.

Serial interval

Definition *SI*

The serial interval (SI) is the **time between disease onset** of successive cases in a chain of transmission.

- Empirical data from infector-infectee pairs
  - Source: cluster data
  - Epidemiologically very likely connection
  - Date of disease onset of both is known
- Fit gamma, lognormal, Weibull and exponential distribution

SI estimates

- Estimates from chinese data
  - Du et al.: *SI* 3.96 days, standard error 4.75 days (Du et al. (2020))
  - Nishiura et al.: *SI* 4.7 days, standard error 2.9 days (Nishiura, Linton, and Akhmetzhanov (2020))
- Two time periods
  - 23.02.2020 – 01.04.2020 (6 weeks)
  - 06.09.2020 – 17.05.2021 (8 months)
- Transmission pairs
  - 312 during first period
  - 250 during second period
- R-package `fitdistrplus` (Delignette-Muller and Dutang (2015))

SI estimates, 23.02.2020 – 01.04.2020 (Wild type period)

Table 1: Fitted parameters, mean, standard deviation, 95% confidence intervals and AIC.

distribution	parameter	value	mean	mean 95% CI	sd	sd 95% CI	AIC
gamma	α β	2.88 0.65	4.46	4.16–4.76	2.63	2.37–2.90	**1,413.3**
lnorm	μ σ	1.31 0.64	4.54	4.19–4.88	3.21	2.77–3.69	1,425.4
exp	λ	0.22	4.46	3.95–4.96	4.46	3.95–4.96	1,558.3
Weibull	k λ	1.80 5.03	4.47	4.17–4.77	2.57	2.32–2.79	1,421.0

SI estimates, 23.02.2020 – 01.04.2020 (graphical)

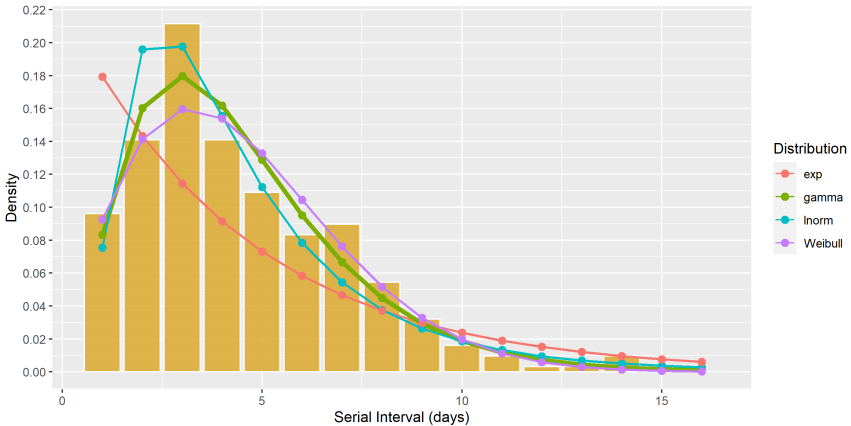


Figure 1: Empirical and fitted distributions of the serial interval of 312 Covid19 transmission pairs observed between 23.02.2020 and 01.04.2020, Austria. The gamma distribution which yields the best fit is highlighted with a thicker line.

SI estimates, 06.09.2020 – 17.05.2021 (Alpha period)

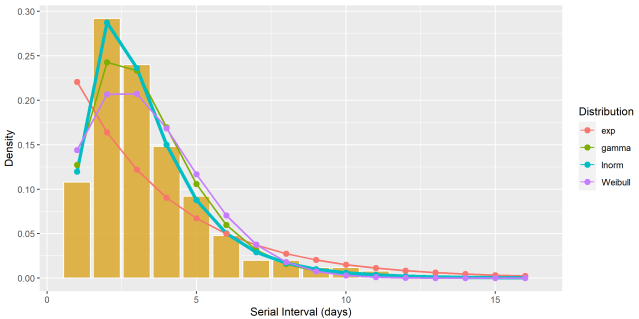


Figure 2: Empirical and fitted distributions of the serial interval of 250 Covid19 transmission pairs observed between 06.09.2020 and 17.05.2021, Austria. The lognormal distribution which yields the best fit is highlighted with a thicker line.

- Mean of lnorm: 3.38 (95% CI: 3.12 – 3.62)
- SD of lnorm: 2.05 (95% CI: 1.74 – 2.37)

Basic and effective reproduction number

Definition R_0

The basic reproduction number is the average number of secondary cases per infectious case in a population where **everyone is susceptible**.

Definition R_{eff}

The effective reproduction number is the average number of secondary cases per infectious case in a population made up of both **susceptible and non-susceptible** hosts.

Basic and effective reproduction number

Definition R_0

The basic reproduction number is the average number of secondary cases per infectious case in a population where **everyone is susceptible**.

Definition R_{eff}

The effective reproduction number is the average number of secondary cases per infectious case in a population made up of both **susceptible and non-susceptible** hosts.

- $R_{eff} > 1 \Rightarrow$ increasing number of cases.
- $R_{eff} = 1 \Rightarrow$ endemic transmission.
- $R_{eff} < 1 \Rightarrow$ decreasing number of cases.
- Implicitly includes effects of interventions.

R_{eff}, Cori Method (Cori et al. (2013))

- $y_t \sim Pois(\lambda_t)$, number of new cases at day t
- w_s, probability that a case generates another case s days after its infection
- s, *generation time* with density $p(s;.)$
- $s \approx SI$
- For $s = 1, \ldots, m$, $w_s = \frac{p(s;.)}{\sum_{i=1}^{m} p(i;.)}$ and $\sum_{i=1}^{m} w_s = 1$
- R_{eff} is the mean number of cases generated by one infected
- Infected at time $t - s$ contribute with a rate of $R_{eff} w_s$ to the number of cases at day t
- cases at times $\{1, \ldots, t - s - 1, t - s + 1, \ldots, t - 1\}$ also contribute to the number of cases at day t
- λ_t denotes the mean number of cases at day t, therefore $\lambda_t = R_{eff} \sum_{s=1}^{t} y_{t-s} w_s$

R_{eff}, Cori Method cont.

- Assumption: R_{eff} for day t is constant over a period of τ days, hence we write $R_{t,\tau}$
- Estimate $R_{t,\tau}$ using Bayesian inference

$$R_{t,\tau} \sim \text{Gamma}\left(a + \sum_{i=t-\tau+1}^{t} y_i, \frac{1}{\frac{1}{b} + \sum_{i=t-\tau+1}^{t} \sum_{s=1}^{i} y_{i-s} w_s}\right)$$

with a, b parameters of the prior gamma distribution

- Mean

$$\hat{R}_{t,\tau} = \frac{a + \sum_{i=t-\tau+1}^{t} y_i}{\frac{1}{b} + \sum_{i=t-\tau+1}^{t} \sum_{s=1}^{i} y_{i-s} w_s}$$

- 95% CI calculated as the 2.5% and the 97.5% percentile of the Gamma distribution
- R-package `EpiEstim`

RKI Method

- Much simpler approach
- y_t, number of incident cases at day t
- τ, number of days included in the estimation of R_{eff} as above
- s, an integer denoting the generation time in full days
- RKI estimate of $R_{t,\tau}$ is calculated as

$$R_{t,\tau} = \frac{\sum_{i=t-\tau+1}^{t} y_t}{\sum_{i=t-\tau+1}^{t} y_{t-s}} = \frac{\bar{y}_t^{\tau}}{\bar{y}_{t-s}^{\tau}},$$

where $\bar{y}_t^{\tau} = \frac{1}{\tau} \sum_{i=t-\tau+1}^{t} y_t$ is the moving average of cases of τ days.

Comparison of methods

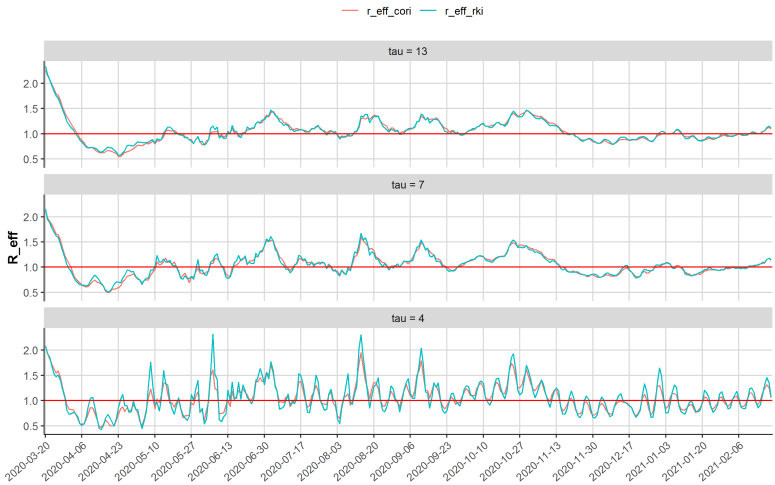


Figure 3: Comparison of the effective reproduction number $R_{t,\tau}$ for the Cori and RKI method for different values of τ.

Imputation of date of onset

- Date of onset is a key variable
- Is often unknown

Imputation of date of onset

- Date of onset is a key variable
- Is often unknown

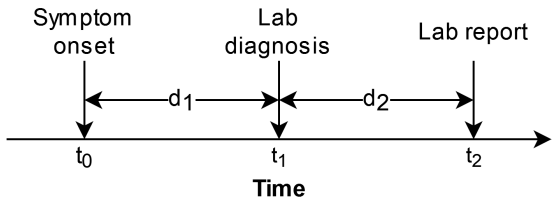


Figure 4: Diagram of the typical chronological order of a case from symptom onset to reporting.

Imputation of date of onset

- Date of onset is a key variable
- Is often unknown

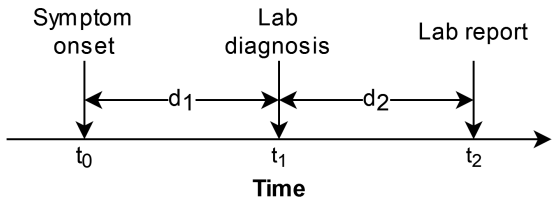


Figure 4: Diagram of the typical chronological order of a case from symptom onset to reporting.

- date of reporting (t_2) vs. date of onset (t_0)
- reporting delay: $t_{diff} := t_2 - t_0$

Imputation of date of onset II

- $t_{diff} \sim \text{Gamma}(\alpha, \beta)$, $\alpha, \beta > 0$.
- PDF: $p(y; \mu, \sigma) = \frac{y^{\frac{1}{\sigma^2}-1} \exp\left(-\frac{y}{\sigma^2 \mu}\right)}{(\sigma^2 \mu)^{\frac{1}{\sigma^2}} \Gamma\left(\frac{1}{\sigma^2}\right)}$ with $\mu = \alpha/\beta$ and $\sigma = 1/\sqrt{\alpha}$
- `gamlss`-model (*flexible generalized additive model for location, scale and shape*) (Stasinopoulos and Rigby (2007)):

$$\eta_{k,cw} = \beta_{k,0} + \beta_{k,1} f(x_{cw}),\ k \in \{\mu, \sigma\}$$

- $f(.)$, smoothing function (cubic splines)
- x_{cw}, calendar week

Results, 2020 week 9 to 22

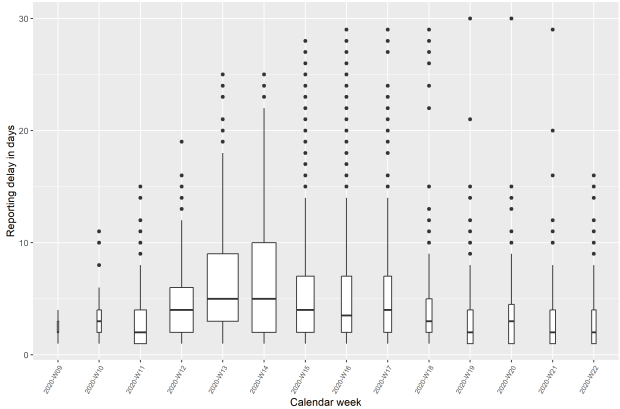


Figure 5: Series of boxplots of the reporting delay t_{diff} by calendar week of cases with known date of report and date of disease onset of SARS-CoV-2, Austria, 2020 week 9 to week 22

- 16,878 cases
- 12,027 (71.3%) with reliable date of onset
- weekly mean t_{diff} ranged from 2.38 – 6.58

Table 2: Parameter estimates of the `gamlss` model for $\hat{\mu}_{cw}$ and $\hat{\sigma}_{cw}$

Parameter	Estimate	Std. Error	p
$\hat{\beta}_{\mu,0}$	1.474	0.023	$< 10^{-3}$
$\hat{\beta}_{\mu,1}$	0.042	0.004	$< 10^{-3}$
$\hat{\beta}_{\sigma,0}$	-0.519	0.018	$< 10^{-3}$
$\hat{\beta}_{\sigma,1}$	0.044	0.003	$< 10^{-3}$

Epicurve and R_{eff}, 2020 week 9 to 22

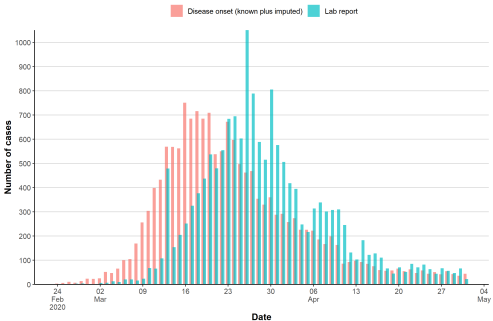


Figure 6: Epicurve by date of disease onset (known and imputed) versus epicurve by date of lab report of SARS-CoV-2, Austria, 24.02.2020 (week 9) to 01.05.2020 (week 18)

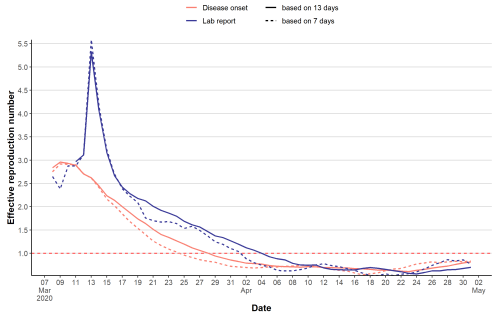


Figure 7: Effective reproduction number based on date of disease onset (known and imputed, red) and based on date of lab report (blue) and based on different values of τ ($\tau = 13$ solid line, $\tau = 7$ days dashed line) of SARS-CoV-2 cases, Austria, 08.03.2020 (week 10) to 01.05.2020 (week 18)

Imputation of variant data

- December 2021 to January 2022: emergence of Omicron Variant (B.1.1.529)
- Has mutation at the 501th segment of the genome: N501Y variant
- We define
  - $x_{n501y} \in \{0, 1, NA\}$, a case has the N501Y mutation
  - $x_{omicron} \in \{0, 1, NA\}$, a case has the Omicron variant
- Two-stage imputation

Imputation of variant data

- December 2021 to January 2022: emergence of Omicron Variant (B.1.1.529)
- Has mutation at the 501th segment of the genome: N501Y variant
- We define
  - $x_{n501y} \in \{0, 1, NA\}$, a case has the N501Y mutation
  - $x_{omicron} \in \{0, 1, NA\}$, a case has the Omicron variant
- Two-stage imputation

A case can enter one of four states:

1. $x_{n501y} = NA \Rightarrow x_{omicron} = NA$
2. $x_{n501y} = 0 \Rightarrow x_{omicron} = 0$
3. $x_{n501y} = 1$ and $x_{omicron} = NA$
4. $x_{n501y} = 1$ and $x_{omicron} \in \{0, 1\}$

Modelling x_{n501y} and $x_{omicron}$

- $x_{n501y} \sim \text{Bernoulli}(r)$
- standardised binomial model using province, calendar week and day of week as explanatory variables:

$$logit(r) = \beta_0 + \beta_1 x_{cw} + \beta_2 x_{prov} + \beta_3 x_{dow} + \varepsilon$$

Modelling x_{n501y} and $x_{omicron}$

- $x_{n501y} \sim \text{Bernoulli}(r)$
- standardised binomial model using province, calendar week and day of week as explanatory variables:

$$logit(r) = \beta_0 + \beta_1 x_{cw} + \beta_2 x_{prov} + \beta_3 x_{dow} + \varepsilon$$

- From above: $x_{n501y} = 0 \Rightarrow x_{omicron} = 0$ and if $x_{n501y} = 1 \Rightarrow x_{omicron} \sim \text{Bernoulli}(s)$
- Analogously

$$logit(s) = \gamma_0 + \gamma_1 x_{cw} + \gamma_2 x_{prov} + \gamma_3 x_{dow} + \varepsilon$$

Imputation of missing data

- Impute missing values of $\hat{x}_{omicron}$ where x_{n501y} is known by randomly imputing binary values based on:

$$
\begin{aligned}
&P(\hat{x}_{omicron} = 1 | x_{n501y} = 1) = \hat{s}, \\
&P(\hat{x}_{omicron} = 0 | x_{n501y} = 1) = 1 - \hat{s}, \\
&P(\hat{x}_{omicron} = 0 | x_{n501y} = 0) = 1 \text{ and} \\
&P(\hat{x}_{omicron} = 1 | x_{n501y} = 0) = 0.
\end{aligned}
$$

Imputation of missing data

- Impute missing values of $\hat{x}_{omicron}$ where x_{n501y} is known by randomly imputing binary values based on:

$$
\begin{aligned}
&P(\hat{x}_{omicron} = 1 | x_{n501y} = 1) = \hat{s}, \\
&P(\hat{x}_{omicron} = 0 | x_{n501y} = 1) = 1 - \hat{s}, \\
&P(\hat{x}_{omicron} = 0 | x_{n501y} = 0) = 1 \text{ and} \\
&P(\hat{x}_{omicron} = 1 | x_{n501y} = 0) = 0.
\end{aligned}
$$

- Randomly impute $\hat{x}_{n501y}$ and $\hat{x}_{omicron}$ for cases with $x_{n501y} = NA$ and $x_{omicron} = NA$ based on:

$$
\begin{aligned}
&P(\hat{x}_{omicron} = 0, \hat{x}_{n501y} = 0) = P(\hat{x}_{omicron} = 0 | \hat{x}_{n501y} = 0) P(\hat{x}_{n501y} = 0) = 1 - \hat{r}, \\
&P(\hat{x}_{omicron} = 0, \hat{x}_{n501y} = 1) = P(\hat{x}_{omicron} = 0 | \hat{x}_{n501y} = 1) P(\hat{x}_{n501y} = 1) = (1 - \hat{s})\hat{r}, \\
&P(\hat{x}_{omicron} = 1, \hat{x}_{n501y} = 1) = P(\hat{x}_{omicron} = 1 | \hat{x}_{n501y} = 1) P(\hat{x}_{n501y} = 1) = \hat{s}\hat{r}.
\end{aligned}
$$

Summary of cases for variant imputation

Table 3: Summary of SARS-CoV-2 cases during the emergence of the Omicron variant, Austria, 22.11.2021 to 14.02.2022

	n	**% of total cases**
Total cases	1,253,256	100.0%
N501Y status and variant known	348,727	27.8%
N501Y status known and variant unknown	1,038	0.1%
N501Y and variant unknown	903,491	72.1%
Other variant known	68,786	5.5%
Omicron variant known	279,941	22.3%
Other variant imputed	154,364	12.3%
Omicron variant imputed	750,165	59.9%
cases with variant data imputed	904,529	72.2%

Model results

Table 4: Binomial model estimates for r (being N501Y) and s (being Omicron), 22.11.2021 to 14.02.2022

	Dependent variable:	
	r	s
	(1)	(2)
kw	1.269*** (0.006)	−0.016 (0.015)
dow2	0.299*** (0.030)	0.197** (0.089)
dow3	0.475*** (0.032)	0.008 (0.088)
dow4	0.574*** (0.032)	0.179** (0.089)
dow5	0.726*** (0.033)	0.212** (0.091)
dow6	0.864*** (0.034)	0.473*** (0.107)
dow7	0.994*** (0.033)	0.466*** (0.108)
blKärnten	−0.493*** (0.062)	
blNiederösterreich	−0.337*** (0.056)	
blOberösterreich	−0.491*** (0.055)	
blSalzburg	0.839*** (0.059)	
blSteiermark	−0.112* (0.062)	
blTirol	0.181*** (0.057)	
blVorarlberg	−0.375*** (0.065)	
blWien	0.302*** (0.053)	
Constant	−124.557*** (0.555)	6.602*** (1.479)
Observations	347,181	281,511
Log Likelihood	−48,455.860	−9,701.072
Akaike Inf. Crit.	96,943.710	19,418.140
Note:		*p<0.1; **p<0.05; ***p<0.01

- Very few (≤ 10) non-Omicron cases among N501Y cases for 4 of 9 provinces
- x_{prov} omitted from model for s
- Almost all variables are significant
  - Sample size effect?

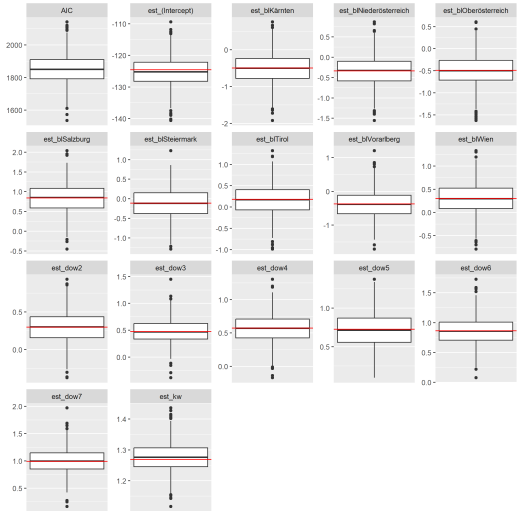
Model results of a subset

Table 5: Subset: Binomial model estimates for r (being N501Y) and s (being Omicron), 22.11.2021 to 14.02.2022

	Dependent variable:	
	r	s
	(1)	(2)
kw	1.276^{***} (0.041)	-0.027 (0.097)
dow2	0.224 (0.232)	-0.496 (0.710)
dow3	0.341 (0.232)	-0.881 (0.681)
dow4	0.029 (0.224)	0.627 (0.914)
dow5	0.597^{**} (0.238)	-0.624 (0.693)
dow6	0.666^{***} (0.245)	-0.903 (0.713)
dow7	1.127^{***} (0.254)	-0.468 (0.772)
blKärnten	-0.988^{**} (0.442)	
blNiederösterreich	-0.528 (0.410)	
blOberösterreich	-0.637 (0.402)	
blSalzburg	0.497 (0.421)	
blSteiermark	0.122 (0.463)	
blTirol	-0.305 (0.412)	
blVorarlberg	-0.802^{*} (0.481)	
blWien	-0.037 (0.389)	
Constant	-124.853^{***} (4.042)	8.266 (9.931)
Observations	6,667	5,445
Log Likelihood	-948.006	-213.260
Akaike Inf. Crit.	1,928.012	442.519
Note:		*p<0.1; **p<0.05; ***p<0.01

- Random sample of 23,658 cases
- Much less significant variables
- Model estimates: full vs. subset?

Model estimates full (red) vs. 1000 samples, model of r



- Median of model estimates are similar to full dataset
- A single sample can yield extreme results
- Standard deviation correlates with sample size (not shown)

Figure 8: Boxplots of model estimates of r (being an N501Y case) based on 1,000 samples of size 23,658, 22.11.2021 to 14.02.2022. Results of the full dataset are shown as red lines.

Why impute this data?

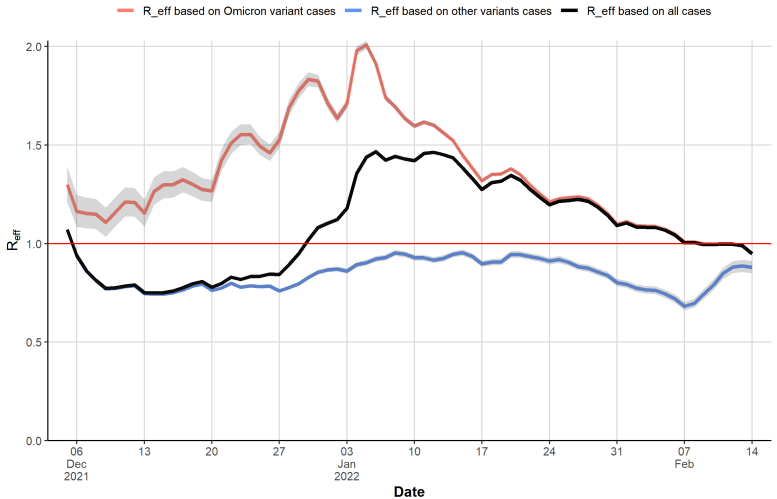


Figure 9: Effective reproduction number based on date of lab diagnosis by (imputed) SARS-CoV-2 variant, Austria, 05.12.2021 to 14.02.2022

Wrap up

- Statistical methods play an important role during outbreaks/pandemics.
- We introduced methods to inform stakeholders and the public about key indicators (SI, R_{eff}).
- Impute missing data to analyse dynamics of new variants.

Other applications:

- Estimation of excess mortality
- Vaccine effectiveness
- Nowcasting and Forecasting

References

Cori, Anne, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. 2013. "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics." *American Journal of Epidemiology* 178 (9): 1505–12. https://doi.org/10.1093/aje/kwt133.

Delignette-Muller, Marie Laure, and Christophe Dutang. 2015. "Fitdistrplus: An R Package for Fitting Distributions." *Journal of Statistical Software* 64 (1): 1–34. https://doi.org/10.18637/jss.v064.i04.

Du, Zhanwei, Xiaoke Xu, Ye Wu, Lin Wang, Benjamin J. Cowling, and Lauren Ancel Meyers. 2020. "Early Release - Serial Interval of COVID-19 Among Publicly Reported Confirmed Cases - Volume 26, Number 6—June 2020 - Emerging Infectious Diseases Journal - CDC." https://doi.org/10.3201/eid2606.200357.

Nishiura, Hiroshi, Natalie M. Linton, and Andrei R. Akhmetzhanov. 2020. "Serial Interval of Novel Coronavirus (2019-nCoV) Infections." *medRxiv*, February, 2020.02.03.20019497. https://doi.org/10.1101/2020.02.03.20019497.

Stasinopoulos, D. Mikis, and Robert A. Rigby. 2007. "Generalized Additive Models for Location Scale and Shape (GAMLSS) in R." *Journal of Statistical Software* 23 (1): 1–46. https://doi.org/10.18637/jss.v023.i07.

Special Thanks to:
Ernst Stadlober
Team at AGES

Special Thanks to:
Ernst Stadlober
Team at AGES

Any questions?

Contact: lukas.richter@ages.at

Backup slides

PDFs of selected distributions

- gamma distribution: $p(x;\alpha,\beta) := \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)}, x>0, \alpha>0, \beta>0.$
- lognormal distribution: $p(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right), x>0, \sigma>0$
- exponential distribution: $p(x;\lambda) = \lambda e^{-\lambda x}, x \geq 0, \lambda > 0.$
- Weibull distribution: $p(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, x \geq 0, k>0, \lambda>0.$

SI estimates, 06.09.2020 – 17.05.2021 (Alpha variant period)

Table 6: Fitted parameters, mean, standard deviation, 95% confidence intervals and AIC.

distribution	parameter	value	mean	mean 95% CI	sd	sd 95% CI	AIC
gamma	α β	3.38 1.00	3.37	3.15–3.60	1.83	1.64–2.03	963.3
lnorm	μ σ	1.06 0.56	3.38	3.12–3.62	2.05	1.74–2.37	**954.2**
exp	λ	0.30	3.37	2.97–3.74	3.37	2.97–3.74	1,109.2
Weibull	k λ	1.84 3.81	3.39	3.15–3.62	1.91	1.73–2.08	986.1

R_{eff}, Cori Method, Bayesian inference I

- As prior for $R_{t,\tau}$ we choose a Gamma distribution and we get:

$$p(R_{t,\tau}; a, b) = p(R_{t,\tau}) := \frac{R_{t,\tau}^{a-1} e^{-R_{t,\tau}/b}}{b^a \Gamma(a)}, R_{t,\tau} > 0, a > 0, b > 0 \tag{1}$$

- The likelihood is given by

$$p(\boldsymbol{y}; R_{t,\tau}) = \prod_{i=t-\tau+1}^{t} (R_{t,\tau} d_i)^{y_i} \exp\left(-R_{t,\tau} d_i\right) \frac{1}{y_i!} \tag{2}$$

with $d_i := \sum_{s=1}^{i} y_{i-s} w_s$

R_{eff}, Cori Method, Bayesian inference II

- Posterior:

$$p(R_{t,\tau}; \boldsymbol{y}) = \frac{p(\boldsymbol{y}; R_{t,\tau}) p(R_{t,\tau})}{\int p(\boldsymbol{y}; R_{t,\tau}) p(R_{t,\tau})\, d\, R_{t,\tau}} \tag{3}$$

- Plug (1) and (2) into (3) and reformulate the expression:

$$p(R_{t,\tau}; \boldsymbol{y}) = R_{t,\tau}^{a+\sum_{i=t-\tau+1}^{t} y_i - 1} \exp\left(-R_{t,\tau}\left(\frac{1}{b} + \sum_{i=t-\tau+1}^{t} d_i\right)\right) k(\boldsymbol{y}, t, a, b),$$

with

$$k(\boldsymbol{y}, t, a, b) = \prod_{i=t-\tau+1}^{t} \frac{d_i^{y_i} c}{y_i! b^a \Gamma(a)}$$

- $\Rightarrow p(R_{t,\tau}; \boldsymbol{y})$ is proportional to

$$p(R_{t,\tau}; \boldsymbol{y}) \propto R_{t,\tau}^{a+\sum_{i=t-\tau+1}^{t} y_i - 1} \exp\left(-R_{t,\tau}\left(\frac{1}{b} + \sum_{i=t-\tau+1}^{t} d_i\right)\right) \frac{\frac{1}{b} + \sum_{i=t-\tau+1}^{t} d_i}{\Gamma(a + \sum_{i=t-\tau+1}^{t} y_i)}$$

R_{eff}, Cori Method, Bayesian inference III

- It follows that

$$R_{t,\tau} \sim \text{Gamma}\left(a + \sum_{i=t-\tau+1}^{t} y_i, \frac{1}{\frac{1}{b} + \sum_{i=t-\tau+1}^{t} d_i}\right) \tag{4}$$

- The estimate of R_{eff} at day t and based on τ days is the mean of $R_{t,\tau}$ and is given as

$$\begin{aligned}\hat{R}_{t,\tau} &= \frac{a + \sum_{i=t-\tau+1}^{t} y_i}{\frac{1}{b} + \sum_{i=t-\tau+1}^{t} d_i} \\ &= \frac{a + \sum_{i=t-\tau+1}^{t} y_i}{\frac{1}{b} + \sum_{i=t-\tau+1}^{t} \sum_{s=1}^{i} y_{i-s} w_s}\end{aligned}$$