Felix Günther[1,2], Andreas Bender[1], Katharina Katz[3], Helmut Küchenhoff[1], Michael Höhle[4]
[1]Statistical Consulting Unit StaBLab, LMU Munich, Germany
[2]Department of Genetic Epidemiology, University of Regensburg, Germany
[3] Bavarian Health and Food Safety Authority, Oberschleißheim, Germany
[4]Department of Mathematics, Stockholm University, Sweden

April 23, 2020

Abstract

To assess the current dynamic of an epidemic it is central to collect information on the daily number of newly diseased cases. This is especially important in real-time surveillance, when one aims at evaluating the effects of interventions on disease spread. Reporting delays between disease onset and case reporting hamper our ability to understand the dynamic of an epidemic when looking at the number of daily reported cases only. Nowcasting can be used to adjust daily case counts for occurred-but-not-yet-reported events. Here, we present a novel application of nowcasting to data on the current COVID-19 pandemic in Bavaria. It is based on a hierarchical Bayesian model that considers changes in the reporting delay distribution associated with the week and weekday of reporting and assumes a smooth epidemic curve. Furthermore, we present a way to estimate the time-dependent case reproduction number $R(t)$ based on predictions of the nowcast. We provide methodological details of the developed approach, illustrate results based on data of the current epidemic, discuss limitations and alternative estimation strategies, and provide code for reproduction or adaption of the nowcasting to data from different regions. Results of the nowcasting approach are reported to the Bavarian health authority and published on a webpage on a daily basis.

-Preliminary Version, submitted to Biometrical Journal-

1 Introduction

Daily reported case numbers of an infectious disease outbreak do not correspond to the actual number of new cases on that day. Due to delays in reporting, the number of newly reported cases and the actual number of new cases can substantially differ. It is the latter, however, that is of central interest when assessing the state and dynamic of an epidemic outbreak. Focusing on the daily number of reported cases hampers our ability to understand current dynamics of the outbreak. This is especially problematic when one wants to assess the effects of political and social interventions. Knowledge of the actual number of daily infections is highly relevant for the current COVID–19 pandemic, where far-reaching political action was taken in order to contain the epidemic outbreak in 2020.

The problem of occurred-but-not-yet-reported cases during outbreaks is well known from the HIV/AIDS outbreak and different statistical approaches have been proposed to handle delayed reporting. A standard reference is Lawless (1994). This approach has recently been used by an der Heiden and Hamouda (2020) for analyzing the current state of COVID-19 in Germany. A more flexible Bayesian approach, which is the basis of the model we use here, has been developed by Höhle and an der Heiden (2014). In the following we will refer to this approach as the *nowcast*.

The basic idea of nowcasting is to estimate the reporting delay based on observations where both, the symptom onset and the reporting date are known. Given the onset delay distribution and the current number of reported cases, we can infer the actual number of cases for a specific day based on the reported number of cases on that day. The resulting, estimated epidemic curve gives a more realistic picture of the current state of the epidemic. Furthermore, the nowcast can facilitate estimation of the time-dependent,

effective reproduction number R_t (Wallinga and Teunis, 2004). There are other approaches including mathematical infection models (SECIR-model) for the estimation of R_t, see e.g. (Khailaie et al., 2020).

Using our approach, we provide nowcast estimates for the COVID–19 pandemic in Bavaria using data from the Bavarian Health and Food Safety Authority (LGL) including the estimation of R_t. The results are updated daily with recent data. In this manuscript, we provide methodological details and show results based on data obtained from the LGL on April, 9th 2020, 10am.

2 Data

We use daily data on reported COVID-19 cases from Bavaria from the mandatory notification data based on the German Infection Protection Act (IfSG). The data is provided by the Bavarian Health and Food Safety Authority (LGL).

To adjust daily new case numbers for reporting delays using nowcasting, we need information on the reporting date of cases and the date of disease onset. There are at least two different ways to define the reporting date of new COVID-19 cases: Firstly, one could use the date at which a new case was registered at a local health authority or, secondly, the date at which the case was reported by the regional health authority (i.e. in our case the LGL). We obtain our data from the regional authority and therefore only the latter approach ensures consistent information on newly reported case numbers for the most current days and no retrospective changes in reported daily case numbers due to delay between the reporting at the local and regional authorities. Since such consistent information is necessary to perform valid nowcasts we build our analyses based on the reporting dates from the LGL.

The local health authorities also provide information on disease onset which stems from a retrospective collection of the day of symptom onset. The daily COVID-19 surveillance data of Bavaria contain, however, about 50%-60% cases with missing information on disease onset. This is partly due to the heavy workload imposed on health authorities during the pandemic, but also because there are several cases with no or only very mild symptoms. Furthermore, note that the date of symptom onset does not necessarily correspond to the infection date due to a preceding incubation time.

3 Methods

3.1 Nowcasting

Due to the many cases with missing disease onset date, we decided to proceed with a two-step approach for nowcasting. Firstly, we impute missing data on disease onset and, secondly, perform the nowcast based on the information on reporting date (available for all cases) and the date of disease onset, which is partly available and partly imputed. Imputing missing symptom onset information implies that we also consider asymptomatic COVID-19 cases in the nowcasting and the projected number of daily new cases (to the part at which they are observed in the official COVID-19 case counts). The rationale is, that this allows to compare the nowcasting results to the daily updated reported case numbers. Furthermore, focusing on symptomatic cases only is problematic since for cases with missing disease onset it is not completely clear whether they are asymptomatic, only asymptomatic at the date of diagnosis, or actually show symptoms but information on symptom onset is missing for other reasons.

Imputation of disease onset: For the imputation, we fit a flexible generalized additive model for location, scale and shape (GAMLSS; Stasinopoulos et al. (2017)), assuming a Weibull-distribution for the delay time $t_d > 0$ between symptom onset and reporting date:

$$t_d \sim WB(\mu, \sigma), \mu > 0, \sigma > 0,$$

where μ and σ are the location and scale parameters of the Weibull distribution with density $f(t_d|\mu,\sigma) = \sigma \cdot \mu \cdot t_d^{(\sigma-1)} \exp(-\mu t_d^{\sigma})$. The same, additive predictor (1) structure was defined for both, μ and σ

$$\eta_j = \beta_{0,j} + \sum_{k=1}^{6} \beta_{k,j} I(x_{weekday} = k) + f_{1,j}(x_{week}) + f_{2,j}(x_{age});\ j \in \{\mu, \sigma\}, \tag{1}$$

however, the estimated effects could differ for the two properties. In (1), parameter $\beta_{0,j}$ is the location/scale specific global intercept and $\beta_{k,j}$ is the effect of the weekday on which the report arrived. Furthermore, $f_{1,j}$ and $f_{2,j}$ are smooth effects of the calendar week and age, respectively, both parameterised via cubic splines.

To estimate the model we use data of all cases for which the disease onset date and thereby t_d is available. Afterwards, we impute the delay time t_d, if missing, by sampling from the fitted, conditional Weibull distribution and derive the missing symptom/disease onset date. No imputation is performed for observations for which the symptom onset date is reported.

Since this imputation induces a missing at random assumption conditioned on the predictors of the GAMLSS imputation model, we perform a sensitivity analysis, where we omit (i) all individuals where the reports say explicitly that they were symptom-free and (ii) all individuals with missing information about symptoms.

Bayesian nowcasting: In the nowcasting step, we follow the approach developed by Höhle and an der Heiden (2014) based on the implementation in the R-package `surveillance` (Salmon et al., 2016).

Let $n_{t,d}$ be the number of cases, which occur on day t and are reported with a delay of d days (case report arrives on day $t+d$). On day $T > t$ (current day), information is available on $N(t,T) = \sum_{d=0}^{T-t} n_{t,d}$ cases that occurred on day t and are reported until day T. The aim of nowcasting is to predict the total number of cases on day t, $N(t,\infty) = \sum_{d=0}^{\infty} n_{t,d}$, based on information available up until the current day T. For identifiability reasons, one defines a maximum relevant delay time of $d = D$ and considers each observation with an observed delay $> D$ as having a delay of D. Here we estimate $N(t,\infty)$ using the Bayesian hierarchical model defined below (cf. Höhle and an der Heiden (2014)):

$$\begin{aligned}
\lambda_t &\sim \text{Prior}(\theta),\\
N(t,\infty)|\lambda_t &\sim \text{Po}(\lambda_t),\\
N(t,T)|N(t,\infty), q_{T-t} &\sim \text{Bin}(N(t,\infty), q_{T-t}),
\end{aligned}$$

where $t = 1,\ldots,T$. The parameter λ_t is the expected value of $N(t,\infty)$, the number of newly diseased cases at day t, which is assumed to follow a Poisson distribution. The number of cases diseased at day t and observed up until day T is assumed to follow a Binomial distribution with $N(t,\infty)$ tries and a success probability of q_{T-t}. This success probability q_{T-t} is governed by a further component of the hierarchical model, the time-varying delay distribution. For this delay distribution we utilize a discrete time hazard model $h_{t,d} = P(delay = d|delay \geq d, W_{t,d})$ as

$$\text{logit}(h_{t,d}) = \gamma_d + W'_{t,d}\eta,\ d = 0,\ldots,D-1;\ h_{t,D} = 1,$$

where $W_{t,d}$ is a vector of timepoint- and delay-specific covariates and η the corresponding regression coefficients. With the survival function of the discrete time hazard model, $S_t(d) = P(delay \geq d|time = t)$, we can derive $q_{T-t} = 1 - S_t(T-t+1) = 1 - \prod_{d=0}^{T-t}(1-h_{t,d})$. Altogether, our hierarchical model jointly estimates the delay distribution as well as the epidemic curve.

In the discrete time hazard model, we use breakpoints of the delay distribution every 14 days and estimate a categorical weekday effect with a common effect for holidays and Sunday, since there are substantial differences in the reported case numbers over the week. Furthermore, we set the maximum relevant delay to $D = 21$.

Two different assumptions are investigated with respect to the prior distribution for the expected number of new cases per day λ_t: the first approach is based on a smooth spline on the TP-basis and the second approach is non-parametric with no smoothing using a log-gamma prior.

To account for the variability from the imputation in the nowcast, it is possible to repeat the imputation and estimation of the nowcast several times and average the estimates of daily new cases.

3.2 Estimation of the time-dependent case reproduction number $R(t)$

Once a depletion of susceptibles occurs during an outbreak of a person-to-person transmitted disease or specific interventions are made, a key parameter to track is the so called effective reproduction number (aka. case reproduction number). This time-varying quantity is defined as follows: Consider a case with

symptom onset on day t - the expected number of secondary cases one such primary case generates will be denoted by $R_e(t)$. The time these secondary cases will show symptoms is governed by the serial-interval distribution, which is defined as the time period between manifestation of symptoms in the primary case to time of symptom manifestation in the secondary case Svensson, Å. (2007).

We estimate the time-dependent case reproduction number by the procedure of Wallinga and Teunis (2004): Consider a case j showing symptoms for the first time on day t_i. The relative likelihood that a case i was infected by j is given by

$$p_{ij} = \frac{g(t_i - t_j)}{\sum_{k \neq i} g(t_i - t_k)},$$

where g is the PMF of the serial-interval distribution. For the serial interval distribution we use a discretized version of the results from Nishiura et al. (2020), which find a log-normal distribution with mean 4.7 days and standard deviation 2.9 as the most suitable fit to data from 28 infector-infectee pairs. An estimate of the effective reproduction number at time t is now given as the average reproduction number of each case j showing first symptoms of the illness on day t:

$$\hat{R}_e(t) = \frac{1}{|j : t_j = t|} \sum_{j:t_j=t} \sum_{i \neq j} p_{ij} \tag{2}$$

We prefer the above $R_e(t)$ estimation method over the method used in an der Heiden and Hamouda (2020), because it is unbiased for our generation time distribution – see the discussion in Höhle (2020). For each imputed dataset, we extract $K = 500$ time series of case counts from the posterior distribution of the nowcast and then estimate $R_e(t)$ as described above for each time series using the R-package `R0` Obadia et al. (2012). Furthermore, each $R_e(t)$ estimation generates $M = 100$ samples from the corresponding sampling distribution of $R_e(t)$. Altogether, we report $\hat{R}_e(t)$ as mean of these $K \cdot M$ samples together with the 2.5% and 97.5% quantiles to form a 95% credibility interval for $R_e(t)$. We estimate $R_e(t)$ for all t s.t. $t + q_g(0.95) \leq T$, where $q_g(0.95)$ is the 95% quantile of the serial interval distribution. This avoids a downward bias in the $R_e(t)$ estimation near "now". Alternatively, one could employ correction methods near T (Cauchemez et al., 2006).

3.3 Implementation

All calculations were performstabed using the `R` language environment (R Core Team, 2020). Nowcasting was performed based on a customized version of the `nowcast` function from the `surveillance` package (Salmon et al., 2016), estimation of $R_e(t)$ based on code of the `R0` package (Obadia et al., 2012).
Code for reproducing our analysis and adapting it to other application scenarios is available at `https://github.com/FelixGuenther/nc_covid19_bavaria`. There, we also provide an artificial dataset based on the observed reporting dates of cases but featuring only artificial information on the age and symptom onset dates of the cases.

4 Results

4.1 Data

We present results based on data obtained from LGL on April 9, 2020, 10am. The data contain information on 29,262 COVID-19 cases which we restrict to 29,246 cases reported after 2020-03-01 since the first 16 COVID-19 cases reported between 2020-01-28 and 2020-02-13 (reported disease onset between 2020-01-23 and 2020-02-03, 3 with missing onset information) concerned a contained outbreak as no further cases were detected upon 27.02.2020, and can therefore be assumed to not contribute to the later disease spread.
Information on disease onset is available for 13,137 cases, but reported disease onset was past the official reporting date for 50 cases and before 2020-01-23 for 16. We set the disease onset date for these cases as missing, yielding 13,071 cases with valid information on disease onset (44.7%). For these, the median delay

between disease onset and reporting was 7 days (25%-quantile: 5, 75%-quantile: 11), Table 1 shows observed delay times over the observation period and reveals a considerable increase in the delay distribution over time.

4.2 Imputation of missing disease onset

For imputation of missing disease onset dates we estimate a Weibull GAMLSS with smooth effects of the reporting week and the cases' age, and a categorical effect of the weekday of report arrival on location and scale. Figure 1 shows the estimated association of the covariates with the median delay. All covariates turned out to be relevant, we find an increase in expected delay time over the reporting weeks, lower reporting delay for older cases and differences over the course of a week. The estimated GAMLSS model is used to impute the date of disease onset for cases with missing onset information.

4.3 Nowcasting

Figure 2 shows the number of daily reported cases and the number of cases with reported disease onset on a certain day over time. Furthermore, we display the estimated new cases from nowcasting using the two different priors for λ_t. We observe a clear difference between the estimated new cases from the nowcast and the daily numbers of reported cases. The induced bias due to the reporting delay is obvious: the estimated new daily cases stabilize from around March 20th on while the reported cases still show a rapid increase. The results of the two different nowcasting models are similar in structure. The 95%-prediction intervals show substantial uncertainty in estimates. Note that we set the current day for the nowcasts to April 8th, 2020, since we only consider days with fully available reporting data. Furthermore, we set a reporting lag between the current date and reported nowcast results of two days due to considerable uncertainty in the nowcasts for dates with very few obervations with reported or imputed disease onset. Figure 3 provides an additional illustration of the nowcast results as daily reported to the LGL. It shows the daily observed and imputed incident cases, as well as the estimated new cases based on the nowcast with TP-spline prior for λ_t.

4.4 Estimation of the time-dependent case reproduction number

Figure 4 depicts the estimated $R_e(t)$ as defined in (2) for the time frame from 24th of February until the 28th of March. This time range is defined by the time of the first secondary case observed in the data and the date of the nowcast minus the number of days it takes for 95% of secondary cases to be observed, which is determined based on 95% quantile of the assumed generation time distribution. According to the estimate, $R_e(t)$ decreased steadily since the beginning of the outbreak and is about $R_e(t) = 1$ since March 22nd, with $R_e(t) = 0.94$ ($CI = (0.89, 0.99)$) on March 28th. However, care is required, if interpreting this result with the timing of interventions, because the $R_e(t)$ estimator is defined forward in time.

5 Discussion

Our analyses show that nowcasting is a valuable real-time tool to gain situational awareness in the middle of an outbreak situation. However, there are important limitations of any nowcasting estimation: (i) we correct for a bias due to delays in case reporting, but provide no correction for possible cases in the population that were not tested. This is a big issue in understanding COVID–19 spread, since there are possibly many undetected cases. Assuming a constant factor of under-reporting we can analyze the dynamics of the outbreak in a more reliable way by our nowcasting method. However, if the proportion of undetected cases varies over time, then the dynamics of the epidemic is not described adequately by our approach. (ii) We model the temporal variation in the delay distribution in a flexible way. However, short term changes, especially in the time close to the end can lead to a bias, because it is particularly hard to distinguish between developments in the epidemic curve and changes in the reporting delay with no or very little data. (iii) Our imputation method includes a missing at random assumption, which could

be violated due to many asymptomatic cases of COVID–19 cases. However, the sensitivity analyses in the appendix (5) show that our results are stable to some variations of this definition.

Comparing our approach to that of an der Heiden and Hamouda (2020) we use a more detailed delay distribution modelling for the nowcast, e.g., including the day of the week in our model, which turned out to be relevant in our data. Furthermore, we observed and modeled a dependence of the delay time on calendar time as part of the nowcast. This was not originally taken into account by an der Heiden and Hamouda (2020). When calculating the effective reproduction number $R(t)$, an der Heiden and Hamouda (2020) used a constant generation time of four days, while our approach includes a more realistic assumption of an individually varying time originating from a lognormal distribution.

The approach to estimate $R(t)$ proposed by Khailaie et al. (2020) includes a complex SECIR-model with many assumptions about the other model parameters, which in part can only be guesstimated from literature sources. Their procedure of estimating $R(t)$ is only partly data driven and mainly relies on cumulative reported cases in the federal states of Germany. Confidence intervals are generated by the variation of the other model parameters. This highlights the problems of the approach: While SECIR- models can be useful for forecasting, its value for real-time estimation of $R(t)$ hinges on it being a realistic model with well calibrated parameter estimated. Instead, we prefer the more statistically driven transmission-tree based estimates, which rely less on model assumptions and more on a statistically sound evaluation of the available data.

Summarizing, we believe that our results give a much more reliable picture of the course of the pandemic than the mostly used time series of reported cases. Since we introduce no correction for cases, which are never detected, our estimated epidemic curve should be related to other data sources, like e.g. hospital admission, ICU admission or death numbers. This aspect highlights the need for collecting and combining many different sources - each bringing challenges of its own. Here, a statistical based approach adequately reflecting uncertainty has a lot to offer.

Acknowledgement *We thank Titus Laska for initial code on the $R(t)$ estimation. An early version of the imputation modelling, but without any methodological extensions of the nowcast, was proposed in the now disused work by Glöckner et al. (2020).*

Conflict of Interest *The authors have declared no conflict of interest.*

Appendix

A.1. Sensitivity analysis

Since the amount of data with missing information on disease/symptom onset is rather high, we perform two sensitivity analyses. The symptom onset date can either be missing because the reporting local health authorities were not able to provide any information or because a case did not experience any symptoms until case reporting. Based on available information, we can distinguish between cases for which COVID-19 symptoms are documented at time of reporting(17,723 (61%); 73% with available onset), cases explicitly without symptoms (2,221, 8%) and cases without any information on symptoms (9,302, 32%). In the first sensitivity analysis we focus only on the cases with reported symptoms. In a second analysis, we exclude all cases which have been explicitly reported to have no symptoms. We show nowcast results based on the two hierarchical Bayesian models from the main analysis and additionally using the approach of Lawless (1994).

Figure 5 indicates that the estimated structure of the epidemic curve is very similar to the main analysis when excluding asymptomatic cases and cases without known symptom status in the sensitivity analyses. Since the sensitivity analyses consider fewer reported cases, the actual estimated number of new cases per day is lower as well. The interpretation regarding the current COVID-19 dynamic stays, however, the same.

Results of the Lawless approach differ from the Bayesian estimation at the end of the observation period. This matches our experience from nowcasting for other 'current' dates, where the Lawless method yielded problematic, unstable results at the end of the nowcasting window. This behavior is especially undesirable, since nowcasts for those days are of central interest when interpreting nowcast results.

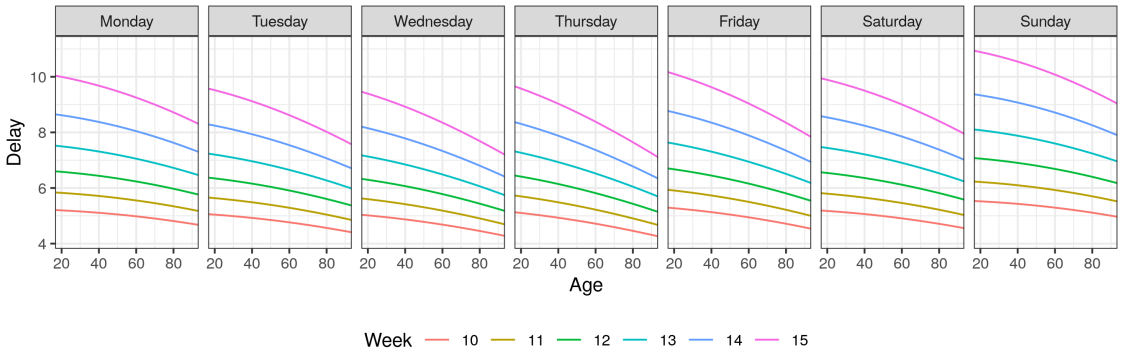

Figure 1: Results of the Weibull GAMLSS imputation model. Shown is the estimated median of the delay time given case-specific covariates (reporting week, weekday of reporting, age).

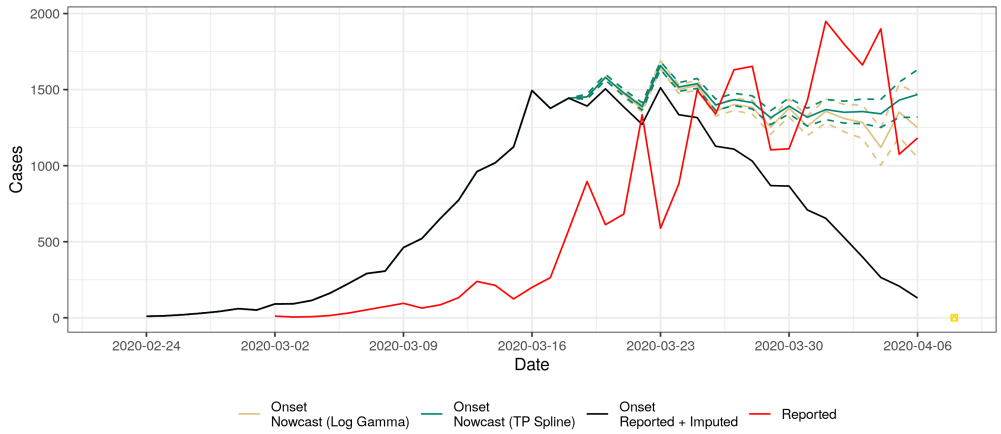

Figure 2: Point estimates + 95%-prediction intervals of the daily number cases with symptom onset on a given day based on two nowcasting models (Green/Yellow). Nowcasts consider changes in delay distribution based on two-week breakpoints and depending on weekdays. Expected number of new cases is modelled based on smooth TP-spline or non-parametric iid log-Gamma prior. Black: number of cases with disease onset on respective date (reported or imputed). Red: newly reported cases by date of report arrival. The current day is April 8th, 2020 and nowcasts are performed up until April, 6th.

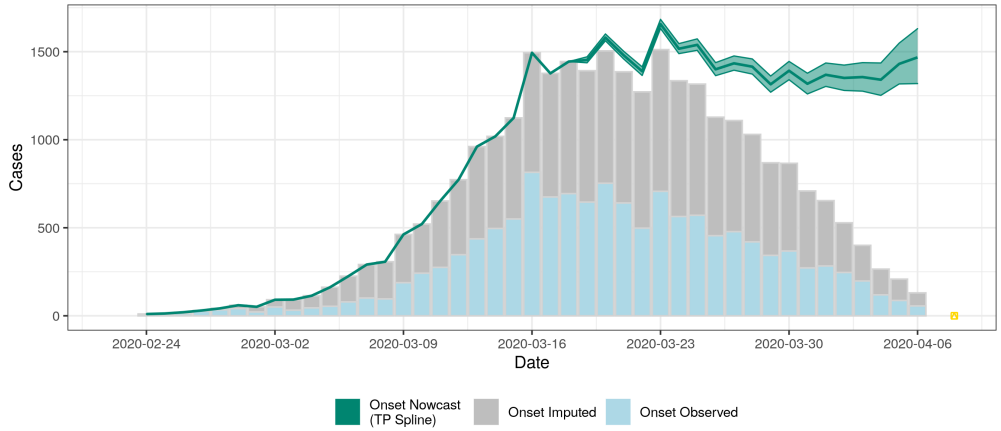

Figure 3: Alternative illustration of nowcast results. The green line and ribbon show the nowcast results and associated 95%-prediction interval of the nowcast with TP-spline modelling for the daily number of expected cases, λ_t. Blue bars show the observed daily number of disease onsets, grey bars the additional imputed cases.

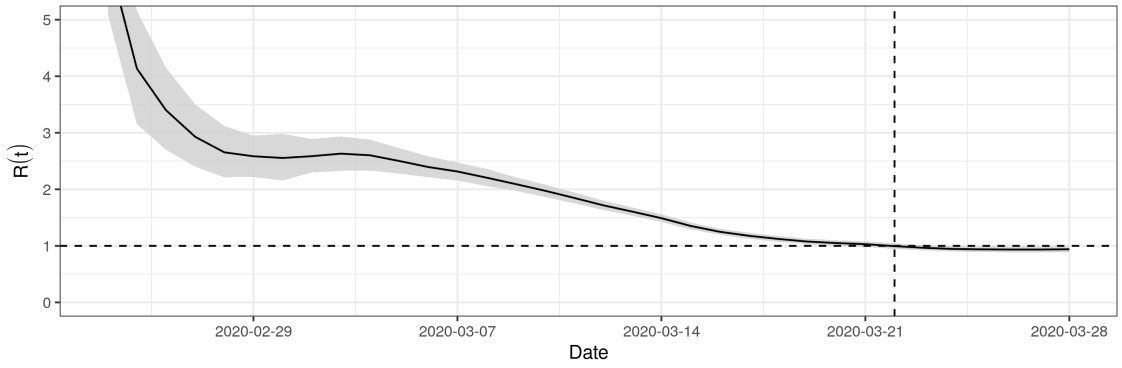

Figure 4: Estimated, time-dependent effective case reproduction number $R_e(t)$.

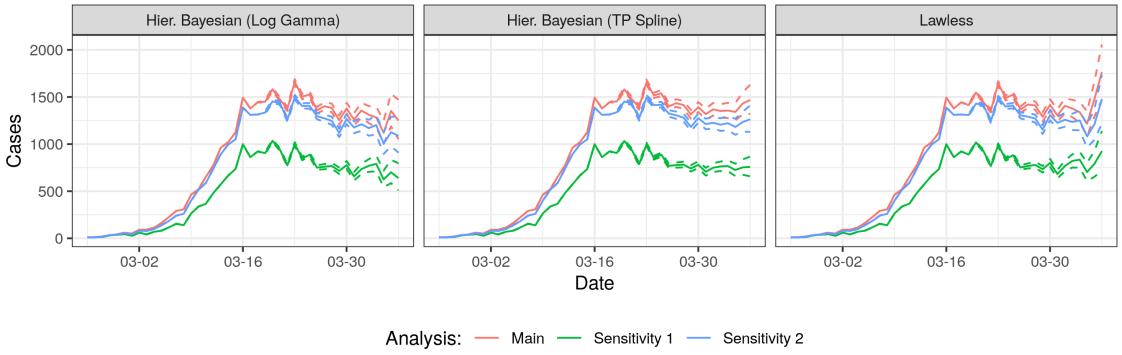

Figure 5: Sensitivity analysis. Three models are included: Hierarchical Bayesian nowcasting using log Gamma or TP-spline prior for λ_t and maximum likelihood approach by Lawless. Three scenarios are presented: original analysis (red), analysis only with cases with reported COVID-19 symptoms (green) and with all cases excluded that are reported as having no symptoms (blue).

Table 1: Week-specific observed number of cases with available information on reporting delay and empirical mean, median and 25%-/75%-quantile of delay distribution.

Week	n	Delay available	% avail.	Mean	Median	25%-Quant.	75%-Quant.
10	69	51	74%	5.5	5	4	7
11	952	451	47%	5.9	5	4	8
12	4566	2010	44%	6.4	6	4	8
13	8693	4216	48%	7.8	7	5	10
14	10926	4826	44%	9.2	8	5	12
15	4040	1517	38%	9.1	7	4	12

References

an der Heiden, M. and Hamouda, O. (2020). Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland – Nowcasting. *Epidemiologisches Bulletin* **17**, 10–15.

Cauchemez, S., Boelle, P. Y., Donnelly, C. A., Ferguson, N. M., Thomas, G., Leung, G. M., Hedley, A. J., Anderson, R. M., and Valleron, A. J. (2006). Real-time estimates in early detection of SARS. *Emerging Infectious Diseases* **12**(1), 110–113.

Glöckner, S., Krause, G., and Höhle, M. (2020). Now-casting the covid-19 epidemic: The use case of japan, march 2020. `https://www.medrxiv.org/content/early/2020/03/23/2020.03.18.20037473`. Accessed: 2020-04-18.

Höhle, M. (2020). Effective reproduction number estimation. `https://staff.math.su.se/hoehle/blog/2020/04/15/effectiveR0.html`. Accessed: 2020-04-16.

Höhle, M. and an der Heiden, M. (2014). Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011. *Biometrics* **70**(4), 993–1002.

Khailaie, S., Mitra, T., Bandyopadhyay, A., Schips1, M., Mascheroni, P., Vanella, P., Lange, B., Binder, S., and Meyer-Hermann, M. (2020). Estimate of the development of the epidemic reproduction number Rt from Coronavirus SARS-CoV-2 case data and implications for political measures based on prognostics. `https://www.medrxiv.org/content/10.1101/2020.04.04.20053637v1`. Accessed: 2020-04-16.

Lawless, J. F. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *The Canadian Journal of Statistics* **22**(1), 15–31.

Nishiura, H., Linton, N. M., and Akhmetzhanov, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases* **93**, 284–286.

Obadia, T., Haneef, R., and Boëlle, P.-Y. (2012). The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making* **12**(1), 147.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Salmon, M., Schumacher, D., and Höhle, M. (2016). Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance. *Journal of Statistical Software* **70**(10), 1–35.

Stasinopoulos, M., Rigby, R., Heller, G., Voudouris, V., and De Bastiani, F. (2017). *Flexible Regression and Smoothing Using GAMLSS in R*. Chapman and Hall/CRC.

Svensson, Å. (2007). A note on generation times in epidemic models. *Mathematical Biosciences* **208**(1), 300–311.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* **160**(6), 509–516.