 doi:10.20944/preprints202009.0628.v1

Exploring COVID-19 daily records of diagnosed cases and fatalities based on simple non-parametric methods

Hans H. Diebner[1,*] **and Nina Timmesfeld**[1]

[1]Ruhr-Universität Bochum, Department of Medical Informatics, Biometry and Epidemiology, 44780 Bochum, Germany
[*]corresponding author: hans.diebner@rub.de

ABSTRACT

Based on comprehensible non-parametric methods, estimates of crucial parameters that characterise the COVID-19 pandemic with a focus on the German epidemic are presented. Where appropriate, the estimates for Germany are compared with the results for six other countries (FR, IT, US, UK, ES, CH) to get an idea of the breadth of applicability and a relational understanding. Thereby, only prevalence data of daily reported new counts of diagnosed cases and fatalities provided by the ECDC are used. Where appropriate, the results are compared with conclusions drawn from using the dataset provided by the RKI. Drawing on uncertain *a priori* knowledge is avoided.

Specifically, we present estimates for the duration from diagnosis to death being 13 days for Germany and about 2 days for Italy as the extremes. Furthermore, based on the knowledge of this time lag between diagnoses and deaths, properly delayed asymptotic as well as instantaneous fatality-case ratios are calculated having superiority compared to the commonly published case-fatality rate. The median of the time series of the instantaneous fatality-case ratio with proper delay of 13-days between cases and deaths for Germany turns out to be 0.024. Asymptotic values are presented for other countries with France ranking highest with a fatality-case ratio of almost 0.2 at its peak. The basic reproduction number, R_0, for Germany is estimated to be between 2.4 and 3.4. The uncertainty stems from uncertain knowledge of the generation time. A delay autocorrelation shows resonances at about 4 days and 7 days, where the latter resonance is at least partially attributable to the sampling process with weekly periodicity. The calculation of the basic reproduction number is based on an evaluation of cumulative numbers of cases yielding time-dependent doubling times as an intermediate step. This allows to infer to the reproduction number during the early phase of onset of the epidemic. In a second approach, the instantaneous basic reproduction number is derived from the incident (counts of new) cases and allows, in contrast to the first version, to infer to the temporal behaviour of the reproduction number during the later epidemic course.

To conclude, by avoiding complicated parametric models we provide insights into basic features of the COVID-19 epidemic in an utmost transparent and comprehensible way.

Keywords: SARS-CoV-2, corona virus, COVID-19, non-parametric model

Date: September 25, 2020

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

1 Introduction

The current (2020) hardly to tackle flood of publications on virological, epidemiological, and sociological aspects of the SARS-CoV-2 corona virus and its related disease COVID-19 [4], along with the concurrent demand by many public health institutions and authorities' for intensifying corresponding research in order to quickly gain deeper understanding of the pandemic, entails a dilemma for researchers. Due to the inevitable lack of overview on existing publications it is almost impossible to ensure that newly published work does not merely add redundancy, thus amplifying the flood.

In spite of this dilemma, the present paper is motivated by the hope that the simplicity of the proposed mathematical methodology applied to data on the incidence of COVID-19 cases leads to meaningful insights. Moreover, it can be generalised and transferred to other epidemics beyond SARS-CoV-2/COVID-19. Thereby, we largely follow the appraisal by S. Jahedi and J. York [15] that complex models as, e.g., dynamical multi-compartment models, are unlikely to be understood by non-experts. Moreover, complex models are usually parametric in nature constructed in order to eventually supply estimates of the involved parameters as, e.g., the basic reproduction number R_0 or the instantaneous effective reproduction number $R(t)$. However, most of these parameters, as $R(t)$, are largely time dependent and are contingent on changing public health policies and social behaviour. Modelling then relies on debatable assumptions on impact and timing of these "soft" criteria and *a priori* guesses of some parameter values.

Arguably, the basic reproduction number, R_0, and the effective reproduction number, $R(t)$ are the most important key figures that classify an epidemic [8]. Originally derived in the context of demographic structured population modelling, where R_0 is defined as the moment of order zero (hence the subscript 0) of the net maternity function, its definition had to be adapted within the scope of infection epidemiology [8, 14]. In the latter context, R_0 is defined for a fully susceptible population at the beginning of an epidemic and refers to the number of secondary cases caused by an index case, whereas $R(t)$ refers to the usually time dependent analogue during the course of the epidemic when the population is no longer fully susceptible.

In Germany, "patient zero", i.e. the index case, was a travelling business person from China visiting a company in Bavaria. The secondary infections caused by this index case have been traced as exactly as possible by Böhmer et al. [3]. We learn from [3] that "patient zero" caused three (detected) secondary infections in Germany and a further (detected) secondary case after returning to China. However, it can be doubted that the strict epidemiological definition of R_0, being an integer number, is meaningful, at least if "index case" really refers to the actual "patient zero." It is not even clear whether the out-of-Germany infection counts. In a sense, "index case" has to be conceived as an "average infectious individual" of the given population, i.e., R_0 can be derived as the expectation value of a Poisson distribution of the number of secondary cases. Therefore, the term "index case" is occasionally found replaced by "one case" [8] or "typical case" [14] in alternative definitions which emphasises the idea of a "representative" case.

Using only secondary infections caused by the index case to estimate R_0 necessitates knowledge about the generation time distribution of the infection, thus creating another insufficiency. In practice, there exist a number of different methods that aim at estimating R_0 from the early approximately exponential phase of an epidemic (see e.g. [8, 23]). An alternative approach to estimate R_0 is related to parameter estimations from fitting differential equation based epidemiological models to incidence data [17], if available.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

Being able to trace an epidemic back to the first infected cases, as for the outbreak in Germany [3], is anything but the usual situation. Most of the freely available data files contain daily counts of newly diagnosed COVID-19 cases as well as recorded deaths as, e.g., the database maintained by ECDC [10]. Thereby, a usually more or less symptom-free proportion of infectees remains undetected and does, therefore, not appear in the dataset. Other datasets (e.g. [16]) additionally contain time series of the number of recovered patients, however, these records do usually not result from rigorously confirmed serological diagnoses but rather from applications of elsewhere estimated average recovery times. Thus, only the records of diagnosed cases and deaths are by and large reliable, at least for most of the countries with an efficient health care system. For a few countries, the reliability of COVID-19 reports and recordings might be questionable.

Occasionally, historical data have later been revised by some countries (cf. annotations in [10]) which may lead to inconsistencies upon reproducing the analysis. However, there are some more serious problems that complicate analysis. As mentioned above, infected individuals with mild or no symptoms are usually not detected. Coverage and frequency of testing heavily depends on local policy as well as on availability and accuracy of diagnostic equipment (cf. [24, 18, 28]) and may change in the course of time leading to a varying ratio of reported counts of cases to unreported numbers of infections. Thus, the recorded diagnosed COVID-19 cases are likely a temporarily non-constant proportion of the number of actual infections. Projected scenarios and forecasting from sophisticated differential equation based epidemiological models (e.g. SIR [20], SEIR [22], fractional SIR models [29] and other [12]) are based on vague assumptions concerning the policy-dependent hence time-dependent factor that scales observed cases up to the real number of infected individuals. Even more vague are modes of translating contact restrictions and other implementations of measures taken to mitigate the impact of the pandemic into the mathematical modelling framework [30, 26].

The approach presented here refrains from discussing complex parametric models and avoids doubtful assumptions on the impact of policies. It goes without saying that large-scale cross-sectional epidemiological studies (for paradigmatic small-scale studies see [28, 2]) are needed to get a reliable quantification of all relevant parameters required for a significant assessment of the pandemic. Meanwhile, the present work provides a comprehensible non-parametric exploration of the existing records of counts of diagnosed cases and fatality events. In brief, we reveal some interesting hallmarks and estimate crucial parameters of the COVID-19 pandemic from the "naked" incidence data without making questionable assumptions which are not directly supported by the dataset. To be specific, we supply country-specific estimates of fatality-case ratios (often confusingly denoted as case-fatality rates) as well as estimates of the average duration from diagnosis to death. In addition, the value of the reproduction number is estimated based on two different types of approximations. Thereby, the presented analysis is restricted to a few countries (Germany, France, Italy, Spain, Switzerland, UK, USA) which are compared to the worldwide situation. Some details are highlighted for Germany. Due to its manageable complexity, the proposed calculations are easily portable not only to datasets of other countries involved in the COVID-19 pandemic but also generally to other epidemic incidence data.

With the proposals of two different functions for the estimation of the time-dependent, i.e., instantaneous reproduction number, we contribute to the currently intense discussion of this important parameter. The first version is defined as a function of the counts of new cases, whereas the second version is a function of cumulative case numbers. This renders the first version as a reliable method to estimate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

effective reproduction numbers when they are close to 1, thus, supports decision making with respect to lowering or strengthening contact restrictions. The second version, due to its smoother (less noisy) time course is well-suited to estimate the basic reproduction number R_0 and the effective reproduction number $R(t)$ for the onset phase of the epidemic.

2 Methods

2.1 Observational Data

In this work, data on the geographic distribution of COVID-19 cases worldwide are used which are made available online by the European Centre for Disease Prevention and Control (ECDC) [10]. The data file contains daily counts of newly diagnosed COVID-19 cases and deaths, stratified by country. The last evaluation date used in this work is September 08, 2020. In addition, incidence time series data for the German epidemic provided by the German Robert-Koch-Institute (RKI) [25] are used to contrast the results, where appropriate. Throughout the article, if not explicitly stated otherwise, the results from analysis refer to the ECDC data.

2.2 Mathematical and Statistical Modelling

2.2.1 Asymptotic and Instantaneous Fatality-Case Ratios

Time within the dataset refers to calendar time with a one per day $(1/d)$ sampling frequency. Therefore, in the following, t refers to a discrete time variable with a spacing of $1d$. To simplify mathematical notation, $t = 0$ refers to the date of first observation and subsequent time points are denoted as $t = 0, 1, \ldots, T$ with $t = T$ being the current or final observation time. However, for an intuitive comprehension of the time scales, the time-axis labels of plots are given in calendar date. The number of newly diagnosed cases at date (time point) t are denoted as $cases(t)$, whereas the cumulative sum of cases up to date t is denoted as $cumCases(t)$ with

$$cumCases(t) = \sum_{i=0}^{t} cases(i). \tag{1}$$

Analogously, the number of daily newly recorded fatalities at date t will be denoted as $deaths(t)$ and the total number of registered deaths up to time point t is denoted by $cumDeaths(t)$.

Since the definition of a "rate" as having the dimension $\frac{1}{t}$ is notoriously disregarded within the community of epidemiologists due to historically established conventions (which particularly holds for the notion of "death rate" and decided Streeck et al. [28] to a lengthy rectification to avoid hostilities), we here explicitly introduce definitions of fatality-measures, which are being applied to the COVID-19 data:

1. Delay$-\Delta t$ asymptotic fatality-case ratio:

$$AFCR_{\Delta t}(t) = \frac{cumDeaths(t)}{cumCases(t-\Delta t)} \qquad \forall\, t \geq \Delta t \tag{2}$$

2. Delay$-\Delta t$ instantaneous fatality-case ratio:

$$IFCR_{\Delta t}(t) = \frac{deaths(t)}{cases(t-\Delta t)} \qquad \forall\, t \geq \Delta t. \tag{3}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

The delay time Δt represents a shift between the two time series $cases(t)$ and $deaths(t)$. Choosing Δt to be the mean duration from diagnosis to deaths is expected to yield the most reliable fatality-case ratio. Confer the following section for a proper optimisation procedure. At the end of the pandemic, formally for $t \to \infty$, $AFCR_{\Delta t}(t \to \infty)$ becomes independent of Δt and converges, at least in the ideal case, to a value that corresponds to what is frequently called case-fatality rate. In real life applications (e.g. in cross-sectional studies like [28]), case-fatality rates are often estimated before the epidemics come to a halt and represent, therefore, only interim values $AFCR_0(t)$ at time t using delay $\Delta t = 0$. The choice of $\Delta t = 0$ can lead to misleading results when the case-fatality rate is estimated at an early stage of the epidemic due to the likely finite survival time $\Delta t > 0$. An extreme example would be an early calculation of $AFCR_0(t)$ yielding zero when the first cases have already been diagnosed up to time point t but no fatality has been reported up to that date.

Of note, $AFCR_{\Delta t}(t)$, even for $t \to \infty$, is not a universal classifier of a pandemic. At best, it classifies the pandemic contingent on particular local health care conditions and policies. It is particularly important in the context of $COVID-19$ and should therefore be emphasised that $AFCR$ (or case-fatality rate) is different from the so-called infection-fatality rate, since $AFCR$ is contingent on testing coverage (cf. [28]), as mentioned in the introduction. The same holds, of course, for $IFCR$. Assuming that the reported fatality events have previously been also reported as diagnosed cases, the fatality-case ratio can be conceived as a proportion of cases that die. Therefore, we occasionally switch to report corresponding percentages.

2.2.2 Diagnosis-to-Death Duration via Maximum Correlation Between Deaths and Time-Delayed Cases

In order to estimate the duration from time of diagnosis to time of death, we introduce the simple approach of maximising Pearson's correlation coefficient of the two time series $deaths(t)$ and $cases(t-\Delta t)$ $(t = \Delta t, \ldots, T)$ as a function of delay time Δt or, alternatively, of $\ln(cumDeaths(t))$ and $\ln(cumCases(t-\Delta t))$. Whether the time lag between $deaths(t)$ and $cases(t)$, i.e. the value of Δt that optimises the delay-time correlation, yields a good approximation to the average diagnosis-to-death duration as estimated from a follow-up of individual cases until their deaths, crucially depends on the presence of a salient temporal pattern in the $cases(t)$ time series which induces a similar time shifted pattern in the $deaths(t)$ time series. In the worst case of a homogeneous time series without epidemic ruptures, the time-delay correlation might be insensitive to detect the diagnosis-to-death duration. In the following, we assume that the proposed method yields an acceptable approximation to diagnosis-to-death duration.

The logarithms for the cumulative data are necessary to scale data to a evaluable range. As a heuristic way to construct confidence intervals for the estimated diagnosis-to-death durations we use the Steiger test of the difference between two independent correlations [27]. Pairwise comparisons of any correlation with the maximum correlation yields a series of p-values. All delays whose correlation coefficient does not significantly differ from the maximum correlation coefficient are defined to lie within the confidence interval of the optimal estimate. Alternatively, the mutual information measure applied to the two time series could be used. However, Shannon entropy and the related Kullback-Leibler divergence which serve as basis for the mutual information tend to relatively lesser discriminate small differences in favour of discriminating larger differences of the two time series (cf. [7]).

2.2.3 Generation Time via Delay-Time Autocorrelation of Cases and Deaths, Respectively

Suppose that t_g is the mean generation time of the SARS-CoV-2 virus. Cases diagnosed at time t should then create a second generation of cases at time point $t + t_g$. It might therefore be worth checking the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

incidence time courses for time-delayed autocorrelations, $C(\Delta t)$. Trivially, the non-delayed autocorrelation should be $C(0) = 1$. For small delays Δt, correlation $C(\Delta t)$ should decline until Δt approaches the generation time $\Delta t = t_g$. However, a plateau or a local maximum of $C(\Delta t)$ around $\Delta t = t_g$ is possible only for non-homogeneously distributed cases and if the variance of the generation time is relatively small. In other words, if the incidence peaks at a given point in time t, e.g. due to a singular event like a mass infections at a large party, a subsequent (damped and widened) peak should be detectable at time point $t + t_g$.

Unfortunately, a non-homogeneity in the data may also arise due to systematic delays in the diagnostic process (e.g. less tests at weekend days) and delays in reporting the data: the "weekend effect". A possible escape from the "weekend effect" could be the usage of deaths records instead of cases. However, by all means, the confounding "weekend effect" has to be kept in mind when evaluating delay-time autocorrelations.

Based on an estimate for t_g, the ratio

$$R(t) = cases(t + t_g)/cases(t) \tag{4}$$

intuitively yields a first rough estimate for the time-dependent effective reproduction ratio. Of course, the next generation of infections is in reality not created all at once after one generation time has passed, i.e., this calculation should be conceived as an orientation. Equation 4 as an approximation to $R(t)$ can additionally be justified by assuming the counts of cases to be Poisson variates. Then, the likelihood that $cases(t - \Delta t)$ counts produce $r_{\Delta t} \cdot cases(t - \Delta t)$ counts Δt days later is given by

$$L = \frac{\left(\sum_{\Delta t=1}^{t} r_{\Delta t} cases(t - \Delta t)\right)^{cases(t)}}{cases(t)!} \mathrm{e}^{-\sum_{\Delta t=1}^{t} r_{\Delta t} cases(t-\Delta t)}. \tag{5}$$

Reducing the distribution of delay-specific contributions $r_{\Delta t}$ to the reproduction $R(t)$ at time point t to a single non-zero value for $\Delta t = t_g$ yields eq. 4 after maximising the likelihood. It is finally worth of note that an estimate of $R(t)$ according to eq. 4 does not depend on the true number of infected individuals as long as the ratio of unreported to diagnosed cases is constant over time.

2.2.4 Piecewise Exponential Growth and the Basic Reproduction Number

For a given time interval $(t, t + \Delta t)$, the epidemic growth can be approximated by an exponential growth

$$cumCases(t + \Delta t) = cumCases(t) \cdot \mathrm{e}^{\lambda_{\Delta t}(t)\Delta t}. \tag{6}$$

The time dependent rate of infection is then given by

$$\lambda_{\Delta t}(t) = \frac{1}{\Delta t} \left[\ln\left(cumCases(t + \Delta t)\right) - \ln\left(cumCases(t)\right)\right]. \tag{7}$$

Rather then $\lambda_{\Delta t}(t)$, the doubling time $t_d(t)$ of an epidemic phase is frequently discussed in the literature [8], which is simply $t_d(t) = \frac{\ln(2)}{\lambda_{\Delta t}(t)}$ for a given interval length Δt.

A well-known approximation to the basic reproduction number R_0 [8] is given by

$$R_0 = 1 + \frac{D \ln(2)}{t_d}, \tag{8}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

with D being the duration of infection, or, more precisely, the duration of infectiousness. In this case, t_d should be the doubling time of the early onset phase of the epidemic. However, we use this formula to estimate an effective time-dependent reproduction number

$$R(t) = 1 + \frac{D\ln(2)}{t_d(t)} = 1 + D\lambda_{\Delta t}(t). \tag{9}$$

Of note, the reproduction number does not determine the duration of an epidemic. Rather, the duration is scaled via the duration of an infected individual being infectious, D. So far, when only using the reported incidence data of COVID-19, the magnitude of D is unknown. We suppose, however, D to be in the same order of magnitude as the generation time, if not identical.

It must be stressed at this point, that eq. 9 strictly holds only at the beginning of the epidemic. Using eq. 9, $R(t)$ has a lower bound of 1, thus, as soon as the doubling time approaches very large values, the approximation eq. 9 for $R(t)$ no longer holds.

3 Results

3.1 Fatality-Case Ratios World-Wide and for 7 Selected Countries

Figure 1 gives a first impression of the world-wide and a few country-specific time courses of cumulative cases and deaths. The time series have been normalised to the world population size (fig. 1A) or the corresponding country population sizes (fig. 1C), respectively. To date (8$^{\text{th}}$ September 2020), the world-wide proportion of diagnosed (reported) cumulative COVID-19 cases reached 0.344% of the world population size and roughly 0.011% deaths (fig. 1A), which corresponds to 3.2% of the diagnosed cases. The corresponding world-wide Delay-0 asymptotic fatality-case ratio $AFCR_0(t)$ (frequently denoted case-fatality rate in the literature) time course is shown in fig. 1B. The time average amounts to 0.043. However, a considerable drift can be observed. Whereas the initial variation during January might be explained as fluctuation due to small numbers of cases and deaths, the drift from February on appears to be systematic. The outbreak started in China and spread with different delays to other countries, which might at least partially play a role for the drift, particularly because countries that joined in later had different policies of testing on social contact restrictions. The enormous rise of mortality until roughly mid May is perhaps due to overwhelmed health care systems. The subsequent decline, to the contrary, is likely due to the increasing frequency of testing for SARS-CoV-2 infections.

A glance onto figs. 1C and figs. 1D confirms that different countries contributed with different relative numbers of cases and deaths to the pandemic. The current cumulative number of cases of the United States reached 2% and the number of deaths 0.057% of the population size. A comparatively low incidence (of registered cases!) can be observed for Germany and Spain. However, while Germany, the USA, and Switzerland each have currently (8$^{\text{th}}$ September 2020) a moderate case-fatality rate below 0.05, France ranks highest with a median value about 0.15, however, it started to decline with the beginning of June 2020. We refrain from going into depth with interpretations, however, an obvious explanation is the relatively low number of tests performed per 1000 inhabitants in France, as has been reported, e.g., by the OECD [24]. It should also be mentioned that COVID-19 mortality is age-related. Thus, countries with a correspondingly age-structured demography like Italy with one of the oldest populations in the world, are perhaps particularly vulnerable to COVID-19 morbidity and mortality [30, 9].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

The sigmoid shape of the curves of the fatality-case ratios is striking. For many countries (including those not shown), the curve starts with fluctuations around a moderate value, followed by a systematic increase to eventually decline towards the end of the curve. We already discussed the impact of testing coverage. However, there is a further crucial aspect that has been neglected so far. Diagnosed individuals with a fatal course die with a certain delay after diagnosis. Therefore, shortly after the first cases have been diagnosed, the fatality curve starts at zero until the first deaths occur. Therefore, we expect that the two curves, $cumCases(t)$ and $cumDeaths(t)$ are shifted against each other by some delay Δt such that the ratio will eventually become constant over time for a proper choice of the delay.

3.2 Diagnosis-to-Death Duration for Germany

Figure 2 shows the result of a delay-dependent correlation analysis applied to the two time series $cumCases(t-\Delta t)$ and $cumDeaths(t)$ with varying delay Δt for the German COVID-19 data. The first panel fig. 2A shows scatter diagrams for logarithmised cumulative deaths, $\ln(cumDeaths(t))$, versus time delayed logarithmised cumulative cases, $(\ln(cumCases(t-delay))$, for a series of 16 subsequent delays $\Delta t = 0, 1, \ldots, 15$. In addition, for each delay, the fitted line resulting from a linear regression is shown along with the values of the corresponding correlation coefficients. For delay $\Delta t = 13d$ the scatter diagram transforms into an almost perfect straight line resulting in a perfect correlation coefficient that assumes 0.998. The question of whether the derived maximum correlation depends on the final observation time, T, i.e. on the lengths of the time series, is addressed in fig. 2B. It can be concluded from fig. 2B that for $T - t_0 > 100d$ a delay of $\Delta t = 13d$ constantly turns out to yield the maximum correlation, however, the curve nearly coincides with the correlation time course for a delay of $\Delta t = 12d$.

The delay-0 asymptotic fatality-case ratio according to eq. 2 is depicted in panel fig. 2C along with the time average (blue line) and median (red line). Finally, fig. 2D shows the delay-13 asymptotic fatality-case ratio along with time average (0.043, blue) and median (0.047, red) corresponding to the optimal delay of $\Delta t = 13d$.

The simple time-delay correlation leads to a convincing estimate for the diagnosis-to-death duration, confirmed by comparing panels fig. 2C and 2D. Early after the outbreak in Germany, the delay-13 asymptotic fatality-case ratio exhibits fluctuations due to rather low counts of deaths in the beginning. From May 2020 on, the ratio remains approximately constant over time suggesting a more or less constant testing coverage, thus constant ratio of undetected to diagnosed case numbers. One of the shortcomings of this "quick-and-dirty" approach is the lack of well-defined information on the variance of diagnosis-to-death duration. However, a heuristic indicator is given by the differences of the delay-specific correlation coefficients around the maximum, which can be tested against the nullhypothesis of no difference using the so called Steiger test [27]. Table 1 lists the estimated correlation coefficients for all delays Δt and the p-values resulting from testing the nullhypotheses of vanishing differences of any one correlation coefficient to the maximum coefficient, in this case that one for delay $\Delta t = 13d$. We conclude from the adjusted p-values that delays $\Delta t = 11d$ and $\Delta t = 14d$ can be conceived as the limits of a confidence interval for the estimated diagnosis-to-death duration of $\Delta t = 13d$. Another approach would include a weighted sum over several delays of the delayed $cumCases(t-\Delta t)$ in the denominator of eq. 2 or other techniques. Following Loy et al. [19], we here trust "the power of our eyes" together with the plausibility provided by the outcome of the Steiger test.

Of note, the optimal delay for a maximum correlation between $\ln(cumDeaths(t))$ and $\ln(cumCases(t-\Delta t))$ on the world-wide scale turns out to be zero. The heuristic confidence in-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

terval based on the Steiger test stretches to $\Delta t = 4d$. However, on this world-wide level, the incidence curves are much too heterogeneous to allow for reliable conclusions on the diagnosis-to-death duration. Presumably, the discrimination of the impact of different delays is hampered by the huge numbers of counts, given the pronounced heterogeneity, thus, excessive dispersion.

In the following, the procedure of maximising correlation is applied to the German incidence time series. We expect a greater power of discriminating the delays since the application to the cumulative counts has a damping effect. Figure 3 has an analogue structure as fig. 2 with the incidence data replacing the cumulative incidence.

Panel fig. 3A shows $deaths(t)$ versus $cases(t - delay)$ for a series of 16 subsequent delays $\Delta t = 0, 1, \ldots, 15$ (in days). For each delay, the fitted line resulting from a linear regression is shown along with the values of the corresponding correlation coefficients. For delay $\Delta t = 13d$, the two time series correlate best with the correlation coefficient assuming the value 0.902. Figure 3B shows that time series comprising more that 100 days lead to robust results, i.e., the derived optimal delay is not contingent on the final observation time.

The delay-0 instantaneous fatality-case ratio according to eq. 3 is depicted in panel fig. 3C along with the time average (blue line) and median (red line). Finally, fig. 3D shows the corresponding delay-13 instantaneous fatality-case ratio along with time average (0.041, blue) and median (0.024, red). The application of this maximum correlation variant gives us the same delay as previously estimated for the cumulative counts. As before for the cumulative incidence data, we compare the correlation coefficients by applying Steiger's test. The result is shown in table 2. From the p-values we construct a confidence interval around the estimated delay $\Delta t = 13d$ ranging from $\Delta t = 12d$ to $\Delta t = 14d$.

As expected, the instantaneous fatality-case ratio shows a more pronounced fluctuation when compared to the corresponding asymptotic fatality-case ratio. However, it is the measure of choice when time dependency of the fatality risk is the case in point. The time series of deaths may exhibit its independent fluctuation, however, a hypothetically temporarily constant fatality-case ratio implies that the temporal variation of the time course of deaths follows the fluctuation of the case incidence curve, albeit with some delay. This gives us the rational behind the assumption that maximum correlation applied to incidence data allows for a more sensitive discrimination of delays. As observed for the delay-13 asymptotic fatality-case ratio (fig. 2D), the delay-13 instantaneous fatality-case ratio (fig. 3D) remains approximately constant from May until beginning of July, followed by a marked drop towards a lower but again approximately constant level of the instantaneous fatality-case ratio. The median value assumes 0.024. This striking result leads us to conclude that the ratio of undetected to diagnosed cases dropped early July. In addition, the increasing coverage of tests applied to children most likely changed the age-structure of the (diagnosed) population.

In the following, this scheme is applied to data from a set of selected countries.

3.3 Diagnosis-to-Death Duration for 7 Selected Countries

In this section, a comprehensive summary plot of the diagnosis-to-death durations for seven selected countries is presented and discussed (fig. 4). The same algorithm as discussed for Germany in the previous section is applied to six further countries (France, Italy, Spain, Switzerland, UK, USA). Concretely, for each of the seven countries' incidence data, a series of coefficients for the correlation between $deaths(t)$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

and $cases(t-\Delta t)$ is calculated with delays Δt ranging from 0 to 17 days. The results are depicted in fig. 4 in form of a heat map. Each column of the panel array represents a country as denoted in the top panel labels. The series of 18 delays for each country is displayed in vertical direction as indicated by the right-hand vertical labels. The magnitudes of the correlation coefficients are colour coded. A glance onto the second column confirms the findings from the previous section: The colour saturation peaks for the delay $\Delta t = 13d$ for the German incidence data which lead us to conclude that late individuals survived in the average 13 days after their COVID-19 diagnosis.

For some countries, as e.g. for the USA and the UK, the correlation coefficients remain at a moderate level for all delays. For both countries, a less marked maximum at delay $\Delta t = 13d$ can be observed. The flat distribution of the magnitudes of the correlation coefficients for the States likely reflects the heterogeneity of sampling incidence data (e.g. spatially as well as temporarily non-constant testing coverage). The same holds for the UK. Italy's diagnosis-to-death durations follow a strictly right-skewed distribution which peaks at very small delays. This points to a test procedure where infected persons are diagnosed rather late and only with severe symptoms. The same holds for Spain. There are weak manifestations of multimodal distributions observed for France and Switzerland. A possible explanation could be a second "resonance" for delays twice as long as the average delay. Again, we refrain from going into depth with interpretations. Of note, however, are the short survival times after diagnosis for Italy and Spain, an insight that is confirm with reports on overwhelmed public health authorities and the generation spanning human-to-human social contact behaviour [9, 24, 30]. Also of note, amongst the 7 countries selected, Germany has the longest and at the same time most reliable diagnosis-to-death duration.

3.4 Estimating Generation Time

The result of the delay-time autocorrelation $C(\Delta t)$ of the $cases(t)$ time series, i.e., the correlation between $cases(t)$ and $cases(t-\Delta t)$, for the German data is depicted in the left panel of fig. 5. The curve shows a decent plateau between the delay $\Delta t = 3-4d$ and a peak at $\Delta t = 6d$. A further plateau is visible at $\Delta t = 12d$. While a generation time between 3 and 6 days appears to be plausible, the observed peak at $\Delta t = 6d$ could also be the impact of the weekend effect (e.g. aggregated counts from the weekend on Monday which are not retro-corrected).

Whereas the COVID-19 testing frequency might be substantially lower at weekends, leading to a biased peak of Monday or Tuesday incidence, the occurrence of fatalities should not depend on the weekday. Unfortunately, although there exist claims of assigning occurrences to the correct date, an assurance is not possible. Having said that, the pronounced local maxima of the delay-time autocorrelation $C(\Delta t)$ of the $deaths(t)$ time series at $\Delta t = 7d$ and $\Delta t = 13d$ are striking. The estimation of the preliminary instantaneous "reproduction ratio" for Germany according to eq. 4 with the delay t_g varying between 1 and 9, is depicted in the right panel of fig. 5. A visual inspection of the produced curves clearly shows that delay $t_g = 7d$ leads to the best reduction of noise, pointing to the superiority of $t_g = 7d$. Apparently, during April and May the contact restrictions had been successful since $R(t)$ remains considerably below 1 during this episode. Starting in June, $R(t)$ exceeded again the threshold of 1, which caused an increased instantaneous incidence.

For the French incidence data, local maxima of the correlation function at almost identical time delays can be observed (see left panel of fig. 6), in fact, even more pronounced than for the German data. Again, the noise is maximally reduced for delay $t_g = 7d$ for the preliminary instantaneous "reproduction ratio" for France (cf. right panel of fig. 6). In comparison to the German time course of $R(t)$, the French

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

instantaneous reproduction ratio is noisier and exhibits rather strong occasional bursts even during the moderate epidemic activity from May on.

It goes without saying that we have to be cautious with conclusions. However, if the observed periodicity results from the weekend effect, it entails an urgent need for quality management of data acquisition since a correct assessment of the COVID-19 epidemic data is pressing. To which proportion the generation time and the weekend effect, respectively, contribute to the observed "resonance" in the delay-time correlation remains an open issue. Figure 7 shows the results based on the RKI dataset in full analogy to fig. 5. The rational behind this comparison is the fact that the RKI provides adjusted dates for the diagnosed cases referring to the time point of the onset of symptoms, if available (see [25]). This reduces the "weekend effect" considerably (less pronounced peak in the delay-time autocorrelation) and leads to an overall smoothing of the incidence curve. Again, a delay of $\Delta t = 7d$ yields the best reduction of noise of the corresponding time course $R(t)$. It appears that $R(t)$ slightly drops below 1 for the final observed week, a feature which is somewhat less pronounced for the ECDC-data.

Despite the aforementioned uncertainties, the suggested methodological approach remains noteworthy. Moreover, independently from its cause, the observed periodicity is important for the assessment presented in the following section.

3.5 Time-Dependent Infection Rate and the Effective Reproduction Number

The instantaneous infection rate (eq. 7) of the German epidemic, assumed to be piecewise constant over short time intervals of length Δt days, is depicted in fig. 8A for a series of 9 intervals from 1 though 9 days. Once more, the interval of 7 days appears to be an optimal choice with respect to noise reduction due to the corresponding periodicity of the incidence time series. If the initial phase before March is skipped due to the uncertain estimation resulting from relatively few counts, the peak in early March can be conceived as a good approximation to the initial infection rate of the epidemic. This peak value corresponds to a doubling time of 2 days. Using eq. 8 to calculate an approximate R_0 yields $R_0 = 1 + \frac{D\ln(2)}{t_d} = 1 + \frac{\ln(2)\cdot 7d}{2d} = 3.4$, where we set $D = 7d$ due to our findings above. However, if we use a 4 day generation time instead, motivated by the moderate plateau emerging at $\Delta t = 4$ in the delay-time autocorrelation depicted in the left panel of fig. 5, then R_0 assumes 2.4.

The entire time series $R(t)$ calculated according to eq. 8 for all 9 chosen intervals $\Delta t = 1, 2, \ldots, 9$ are shown in fig. 8B. It is worth to emphasise again that $R(t)$ computed this way is reliable only for values well above 1. Strictly speaking, eq. 8 is an approximation to R_0, i.e. the basic reproduction number defined for the early epidemic phase.

Again, we additionally calculate the instantaneous infection rate (eq. 7) using the RKI dataset for the purpose of comparison. The result is shown in fig. 9. An overall reduced infection rate can be observed, which gives rise to a lower estimate for R_0 with a maximum below 3. We conclude that the estimated parameter values crucially depend on the quality of data curation.

4 Conclusions and Discussion

Although it is a methodological challenge to derive parameter values relevant to understand the pandemic from pure incidence data without pre-knowledge from independent studies on the magnitude of some of the parameters, we attempted to evaluate incidence data without such an *a priori* knowledge. The rational

behind this enterprise was to refrain from questionable assumptions, in particular in the absence of proper studies that supply evidence to such assumptions.

In this article, we largely took a descriptive stance and reported the figures resulting from the estimations as they are without in-depths interpretations. Occasionally, we suggested possible obvious interpretations. For example, conditional on not surviving the infection (i.e., not in the sense of a censored survival analysis), the average time-to-death after diagnosis (or diagnosis-to-death duration) in Germany is about 13 days, whereas in Italy the time-to-death appears to be rather short ($\simeq$ 2 days), albeit with very low confidence. This could be due to different health system conditions for the two countries being worse for Italy. An even more pessimistic interpretation were to assume the diagnosis to be contingent upon death for some cases. A better understanding likely results from including demographic conditions [9, 30, 1], which was however not considered here.

Amongst the 7 exemplarily analysed countries, France showed the highest fatality-case ratio temporarily being close to 0.2, followed by the UK and Italy having a fatality-case ratio which peak at about 0.15, and the corresponding value in Spain temporarily assumed about 0.12. The United States, Germany, and Switzerland have values close to the world-wide average of about 0.032. Besides the country-specific conditions of health care, another obvious interpretation is the difference in testing coverage, which might explain the huge fatality-case ratio in France who started with mass tests not before May 2020 [24]. Of note, country-specific differences in COVID-19 mortality are not exclusively explainable by differing test frequencies as suggested by the so called excess mortality. Excess deaths, i.e. fatalities which add to the long term temporal average related to the pandemic, are less frequent in Germany compared to the other 6 countries (cf. [5, 11]) included in the present study. Factors are differences in the countries' socio-cultural and demographic background as well as conditions of the health care systems but biological causes like different virus strains and immunological conditions (e.g. due to differing general vaccination status) also cannot be excluded. An in-depth analysis is beyond the scope of the present study.

The calculation of fatality-case ratios using time lags between cumulative cases and deaths turned out to have substantial superiority when being compared to the commonly reported case-fatality rates. This is particularly true for estimates derived well before the epidemic comes to a halt. The correlation-based determination of an optimal time lag gives a good estimate for the average diagnosis-to-death duration and, at the same time, allows for a reliable calculation of the instantaneous fatality-case ratio. In the case of a constant testing coverage, i.e., constant ratio of undetected to diagnosed cases, the delayed fatality-case ratio should also be constant over time, which is what we observed for the German epidemic data. For Germany, deaths appear 13 days delayed with respect to the dates of diagnoses. Compared to 6 other investigated countries, the German diagnosis-to-death duration turned out to be the longest. The shortest duration of less than 3 days had to be ascribed to Italy. The median asymptotic delay$-$13 fatality-case ratio for Germany calculated as the ratio of the 13 days delayed cumulative fatality time series to the time series of cumulative cases assumes 0.047. Using the instantaneous instead of the cumulative incidence data confirms these findings and, arguably, improves the quality of the estimations because damping effects that result from cumulative summation are avoided. Remarkably, median delay$-$13 instantaneous fatality-case ratio for Germany in this case assumes 0.024. This rather low value, compared to the arithmetic time average of 0.044, points to a non-normal distribution since the mean instantaneous fatality-case ratio should always dominate the corresponding asymptotic value due

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

to $\frac{1}{T-\Delta t}\sum_{t-\Delta t=0}^{T}\frac{deaths(t)}{cases(t-\Delta t)} \leq \frac{cumDeaths(T)}{cumCases(T-\Delta t)}$ (cf. [6]). In fact, the distribution of fatalities is extremely right-skewed due to an increased frequency of zero-events towards the end of the observation period.

Attempts to estimate the generation time based on autocorrelation of both cases as well as deaths time series is hampered by the superposition of a suspected weekly periodicity ("weekend effect"). We observed a weak increase of autocorrelation of the German data for a delay of 3-4 days and a stronger increase for a delay of 6-7 days. For the French data, the 3-day-resonance is considerably more pronounced. The generation time is needed to calculate either the basic or the effective (instantaneous) reproduction number. R. Mikut et al. [21] recommend in their approach of estimating the reproduction number to use filter techniques to reduce the weekly periodicity. Such a strategy, however, risks to filter out epidemic relevant delay effects. In their subsequent calculation of R(t), the authors adopt the generation time from other studies' results.

Here we presented two alternatives for the estimation of the reproduction number $R(t)$. The first version is similar to that one used by Mikut et al. [21], though without filter. Simply put, $R(t)$ is calculated as the ratio of counts of new cases at time t to the number of cases at time-lagged time t minus the generation time t_g. We presented graphs for a series of time lags t_g including the "hot" candidates $t_g \simeq 4d$ and $t_g \simeq 7d$. We observed the following: In the proximity of $R(t) = 1$, the moving average of all time courses are almost identical, i.e., independent from the chosen t_g. The chosen time lag has a considerable effect on noise, though, with minimum noise for $t_g = 7d$. For values of $R(t) \gg 1$, the magnitude of noise prevents from deriving a reliable estimate, independently from the chosen generation time. We cautiously conclude, that $R(t)$ calculated this way gives a sufficiently accurate information on the magnitude of $R(t)$ close to 1, which is of relevance for public health decision making. Thereby, opting to use $t_g = 7d$ acts as noise filter whether this time lag is a real epidemic or a sampling effect, or a mixture of both. An estimate of R_0 based on this approach is questionable, which obviously holds independently from the chosen time lag. In this context, it is worth of note that published estimates of R_0 vary in a range between 2.2 and values well above 5 (see e.g. [17, 21, 13] and citations therein). Khailaie et al. [17] presented a similar analysis of the time course of $R(t)$ with values at the beginning of the epidemic in Germany that even exceed 10 in some federal states. In essence, Khailaie et al. came to the same conclusion that a reliable estimate of the basic reproduction number R_0 is hampered by early interventions into the epidemic, thus, only calculations of $R(t)$ for time points $t \gg 0$ have, for the time being, enough confidence to draw policies on.

The second proposed version to calculate $R(t)$ has been based on an intermediate step of firstly computing the instantaneous rate of infection. In simplified terms, we modelled the growth of cumulative cases by means of piecewise exponentials with exponents – the infection rates – held constant within subsequent time windows of equal length. We learnt from varying the length of time windows that once more an interval of 7 days leads to a relatively smooth curve. The time course of the thus derived instantaneous infection rate translates into a time-dependent reproduction rate based on a well-known formula [8]. In contrast to the first version above, this second way of computing $R(t)$ yields reliable estimates for $R(t)$ well above 1, i.e. particularly during the onset of the epidemic before interventions unfold their impact. Thus, this second version perfectly complements the first version above. However, the still existing uncertainty in choosing proper intervals for the stepwise exponentials entails uncertainty with respect to the correct value of R_0. Having said that, our analysis yields an approximate range $2.4 < R_0 < 3.4$, which, arguably, adds evidence to published similar results.

Of note, the results vary between different published datasets. Specifically, we refrained from using the fatalities recorded by the RKI due to the confusing registration dates. Apparently, the registration date corresponds to the date of the diagnose rather than death, which can be learnt only from discussions in the comment section [25] but not from the instruction legend. Moreover, the first reference date January 15, 2020, as it appears in the RKI dataset, is inconsistent with findings published in [3] who date the earliest possible effective contact back to January 20. However, rigorously assessing the quality of data curation is far from straightforward and beyond the intention behind the presented analysis.

In conclusion, we should like to emphasise that we here backed our arguments in part on heuristics and gained insight by trusting in the "power of our eyes" [19], an approach occasionally called "quick and dirty." Our aim was to refrain from complicated mathematical models, to avoid questionable assumptions behind these complex models and, instead, back our arguments on comprehensible algorithms straightforwardly applied to pure incidence data. We very much hope to enrich the existing discourse on this hazardous COVID-19 pandemic.

References

1. O. Adenubi, O. Adebowale, A. Oloye, N. Bankole, H. Adesokan, O. Fadipe, P. Ayo-Ajayi, and A. Akinloye. Level of knowledge, attitude and perception about COVID-19 pandemic and infection control: A cross-sectional study among veterinarians in Nigeria. *Preprints*, 2020. doi:10.20944/preprints202007.0337.v1.

2. Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra-Walker, James Tedrow, Dona Tversky, Andrew Bogan, Thomas Kupiec, Daniel Eichner, Ribhav Gupta, John Ioannidis, and Jay Bhattacharya. COVID-19 antibody seroprevalence in santa clara county, california. *medRxiv*, 2020. doi:10.1101/2020.04.14.20062463.

3. Merle M Böhmer, Udo Buchholz, Victor M Corman, Martin Hoch, Katharina Katz, Durdica V Marosevic, Stefanie Böhm, Tom Woudenberg, Nikolaus Ackermann, Regina Konrad, Ute Eberle, Bianca Treis, Alexandra Dangel, Katja Bengs, Volker Fingerle, Anja Berger, Stefan Hörmansdorfer, Siegfried Ippisch, Bernd Wicklein, Andreas Grahl, Kirsten Pörtner, Nadine Muller, Nadine Zeitlmann, T Sonia Boender, Wei Cai, Andreas Reich, Maria an der Heiden, Ute Rexroth, Osamah Hamouda, Julia Schneider, Talitha Veith, Barbara Mühlemann, Roman Wölfel, Markus Antwerpen, Mathias Walter, Ulrike Protzer, Bernhard Liebl, Walter Haas, Andreas Sing, Christian Drosten, and Andreas Zapf. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. *The Lancet Infectious Diseases*, 2020. doi:10.1016/S1473-3099(20)30314-5.

4. Jeffrey Brainard. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science Mag News*, 2020. doi:10.1126/science.abc7839.

5. Statistisches Bundesamt. Sterbefallzahlen und Übersterblichkeit. www.destatis.de/DE/Themen/Querschnitt/Corona/Gesellschaft/bevoelkerung-sterbefaelle.html, 2020.

6. E. Czuber. Der Mittelwert eines Quotienten. *Journal für die reine und angewandte Mathematik*, 1920(150):175–179, 01 Jan. 1920. doi:10.1515/crll.1920.150.175.

7. Hans H. Diebner, Anna Kather, Ingo Roeder, and Katja de With. Mathematical basis for the assessment of antibiotic resistance and administrative counter-strategies. *PLOS ONE*, 15(9):e0238692, 2020. doi:10.1371/journal.pone.0238692.

8. K. Dietz. The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research*, 2(1):23–41, 1993. PMID: 8261248. doi:10.1177/096228029300200103.

9. Jennifer Beam Dowd, Liliana Andriano, David M. Brazel, Valentina Rotondi, Per Block, Xuejie Ding, Yan Liu, and Melinda C. Mills. Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences*, 117(18):9696–9698, 2020. URL: https://www.pnas.org/content/117/18/9696, arXiv:https://www.pnas.org/content/117/18/9696.full.pdf, doi:10.1073/pnas.2004911117.

10. ECDC. Download today's data on the geographic distribution of COVID-19 cases worldwide. European Centre for Disease Prevention and Control. Covid-19 database, url = https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide, 2020.

11. EuroMomo. Bulletin, week 38. www.euromomo.eu, 2020.

12. Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. Modelling the COVID-19 epidemic and implementation of population-wide interventions in italy. *Nat Med*, 26:855–860, 2020.

13. Xi He, Eric H. Y. Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y. Wong, Yujuan Guan, Xinghua Tan, Xiaoneng Mo, Yanqing Chen, Baolin Liao, Weilie Chen, Fengyu Hu, Qing Zhang, Mingqiu Zhong, Yanrong Wu, Lingzhai Zhao, Fuchun Zhang, Benjamin J. Cowling, Fang Li, and Gabriel M. Leung. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med*, 26:672–675, 2020. doi:10.1038/s41591-020-0869-5.

14. J. A. P. Heesterbeek and K. Dietz. The concept of R_0 in epidemic theory. *Statistica Neerlandica*, 50(1):89–110, 1996. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1996.tb01482.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9574.1996.tb01482.x, doi:10.1111/j.1467-9574.1996.tb01482.x.

15. Sana Jahedi and James Yorke. When the best pandemic models are the simplest. *medRxiv*, 2020. doi:10.1101/2020.06.23.20132522.

16. JHU CSSE. COVID-19 data repository by the center for systems science and engineering (CSSE) at Johns Hopkins University. github.com repository, 2020. URL: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

17. Sahamoddin Khailaie, Tanmay Mitra, Arnab Bandyopadhyay, Marta Schips, Pietro Mascheroni, Patrizio Vanella, Berit Lange, Sebastian Binder, and Michael Meyer-Hermann. Development of the reproduction number from coronavirus SARS-CoV-2 case data in Germany and implications for political measures. *medRxiv*, 2020. doi:10.1101/2020.04.04.20053637.

18. Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490):489–493, 2020. doi:10.1126/science.abb3221.

19. Adam Loy, Lendie Follett, and Heike Hofmann. Variations of q–q plots: The power of our eyes! *The American Statistician*, 70(2):202–214, 2016. doi:10.1080/00031305.2015.1077728.

20. Benjamin F. Maier and Dirk Brockmann. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, 368(6492):742–746, 2020. doi:10.1126/science.abb4557.

21. Ralf Mikut, Tillmann Mühlpfordt, Markus Reischl, and Veit Hagenmeyer. Schätzung einer zeitabhängigen Reproduktionszahl R für Daten mit einer wöchentlichen Periodizität am Beispiel von SARS-CoV-2-Infektionen und COVID-19. Technical report, Karlsruher Institut für Technologie (KIT), 2020. 46.12.01; LK 01. doi:10.5445/IR/1000119466.

22. F. Nazarimehr, V. Pham, and T. Kapitaniak. Prediction of bifurcations by varying critical parameters of COVID-19. *Nonlinear Dyn*, 2020. doi:10.1007/s11071-020-05749-6.

23. Hiroshi Nishiura. Correcting the actual reproduction number: a simple method to estimate R(0) from early epidemic growth data. *International journal of environmental research and public health*, 7:291–302, 2010. doi:10.3390/ijerph7010291.

24. OECD. Testing for COVID-19: A way to lift confinement restrictions. Organisation for Economic Co-operation and Development (OECD) Online, 2020. URL: https://www.oecd.org/coronavirus/policy-responses/testing-for-covid-19-a-way-to-lift-confinement-restrictions-89756248/.

25. Robert Koch-Institut and Bundesamt für Kartographie und Geodäsie. CSV-Datei mit den aktuellen Covid-19 Infektionen pro Tag (Zeitreihe). https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74, 2020.

26. Mahdiar Sadeghi, James M. Greene, and Eduardo D. Sontag. Universal features of epidemic models under social distancing guidelines. *bioRxiv*, 2020. doi:10.1101/2020.06.21.163931.

27. J.H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87:245–251, 1980. doi:10.1037/0033-2909.87.2.245.

28. Hendrik Streeck, Bianca Schulte, Beate Kuemmerer, Enrico Richter, Tobias Hoeller, Christine Fuhrmann, Eva Bartok, Ramona Dolscheid, Moritz Berger, Lukas Wessendorf, Monika Eschbach-Bludau, Angelika Kellings, Astrid Schwaiger, Martin Coenen, Per Hoffmann, Markus Noethen, Anna-Maria Eis-Huebinger, Martin Exner, Ricarda Schmithausen, Matthias Schmid, and Gunther Hartmann. Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. *medRxiv*, 2020. doi:10.1101/2020.05.04.20090076.

29. Amirhossein Taghvaei, Tryphon T. Georgiou, Larry Norton, and Allen R Tannenbaum. Fractional SIR epidemiological models. *medRxiv*, 2020. doi:10.1101/2020.04.28.20083865.

30. Vitaly Volpert, Malay Banerjee, Alberto d´Onofrio, Tomasz Lipniacki, Sergei Petrovskii, and Viet Chi Tran. Coronavirus - scientific insights and societal aspects. *Math. Model. Nat. Phenom.*, 15:E2, 2020. doi:10.1051/mmnp/2020010.

Author contributions statement

Conceptualization: Hans H. Diebner.
Formal analysis: Hans H. Diebner.
Methodology: Hans H. Diebner.
Supervision: Nina Timmesfeld.
Validation: Nina Timmesfeld.
Visualization: Hans H. Diebner.
Writing – original draft: Hans H. Diebner.
Writing – review & editing: Hans H. Diebner, Nina Timmesfeld.

Additional information

The authors declare no competing interests.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

Tables

delay	corr	p	p_adj
0	0.962	0.000	0.000
1	0.967	0.000	0.000
2	0.971	0.000	0.000
3	0.976	0.000	0.000
4	0.980	0.000	0.000
5	0.984	0.000	0.000
6	0.987	0.000	0.000
7	0.990	0.000	0.000
8	0.993	0.000	0.000
9	0.995	0.000	0.000
10	0.996	0.001	0.016
11	0.997	0.174	1.000
12	0.998	0.882	1.000
13	0.998	1.000	1.000
14	0.998	0.239	1.000
15	0.996	0.002	0.032

Table 1. Comparison of correlation coefficients for the cumulative incidence data: Column 2 contains the estimated correlation coefficients of the two time series $\ln(cumCases(t-\Delta t))$ and $\ln(cumDeaths(t))$ with the corresponding delays Δt in days listed in the first column. The p-values in the third column refer to a test for difference of any given correlation coefficient with the maximum correlation coefficient, in this case that one estimated for delay $\Delta t = 13d$. The last column contains the corresponding Benjamini-Hochberg adjusted p-values. Some p-values assume 0.000 after rounding, thus $p < 0.0005$ in such cases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

delay	corr	p	p_adj
0	0.594	0.000	0.000
1	0.575	0.000	0.000
2	0.576	0.000	0.000
3	0.613	0.000	0.000
4	0.678	0.000	0.000
5	0.745	0.000	0.000
6	0.79	0.000	0.000
7	0.796	0.000	0.000
8	0.784	0.000	0.000
9	0.756	0.000	0.000
10	0.77	0.000	0.000
11	0.83	0.003	0.004
12	0.888	0.471	0.502
13	0.902	1.000	1.000
14	0.88	0.274	0.313
15	0.834	0.004	0.005

Table 2. Comparison of correlation coefficients for the incidence data: Column 2 contains the estimated correlation coefficients of the two time series $cases(t-\Delta t)$ and $deaths(t)$ with the corresponding delays Δt in days listed in the first column. The p-values in the third column refer to a test for difference of any given correlation coefficient with the maximum correlation coefficient, in this case that one estimated for delay $\Delta t = 13d$. The last column contains the corresponding Benjamini-Hochberg adjusted p-values. Some p-values assume 0.000 after rounding, thus $p < 0.0005$ in such cases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

Figures

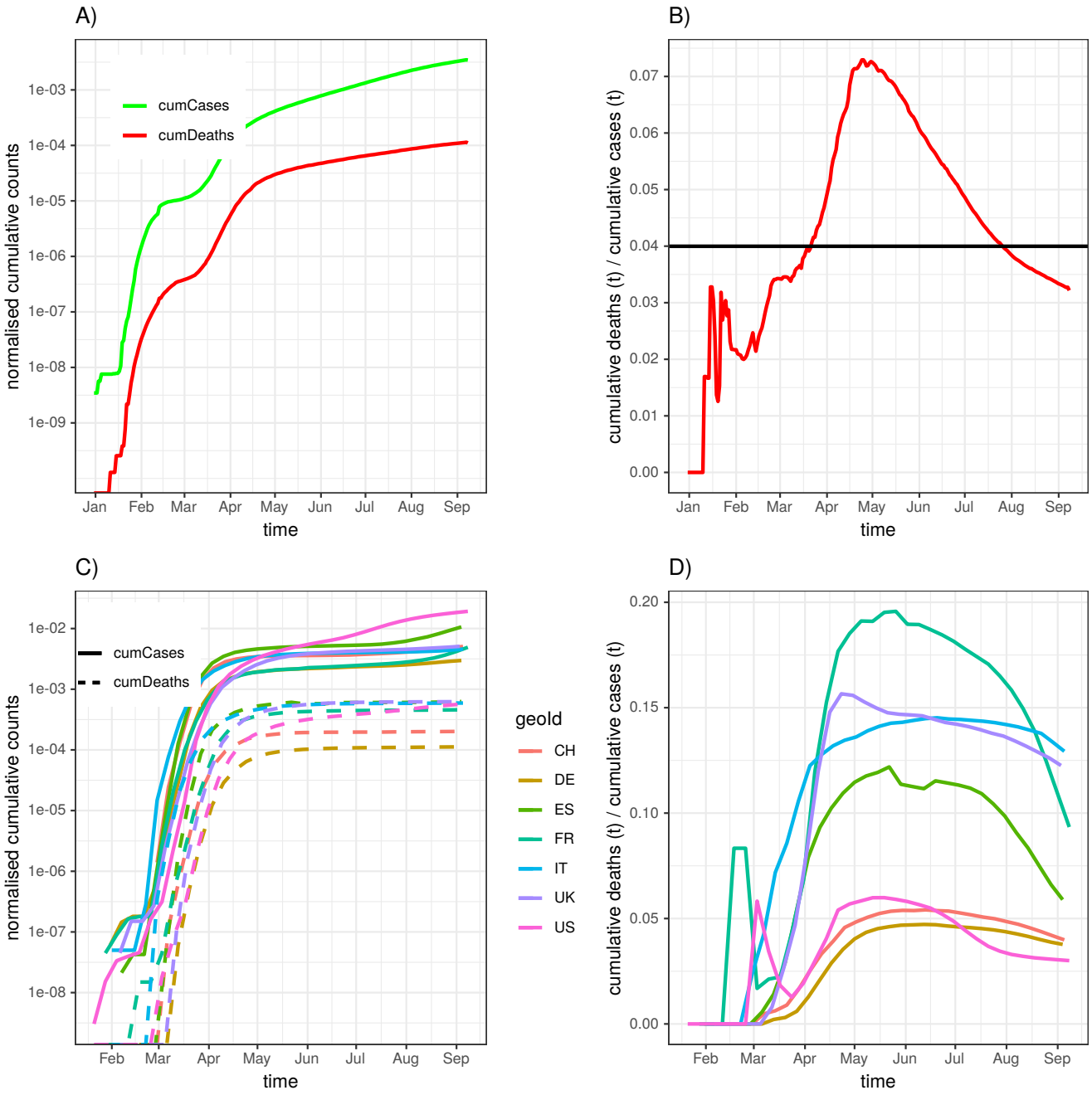

Figure 1. Time courses of cumulative cases, cumulative deaths, and delay−0 asymptotic fatality-case ratios for the entire world and 7 selected countries: A) World-wide cumulative cases and cumulative deaths normalised to the world population. B) World-wide ratio of cumulative deaths to cumulative cases (delay−0 asymptotic fatality-case ratio and the median of 0.04 (black horizontal line). C) Normalised cumulative cases and cumulative deaths of 7 selected countries. D) Delay−0 asymptotic fatality-case ratio for the 7 selected countries.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

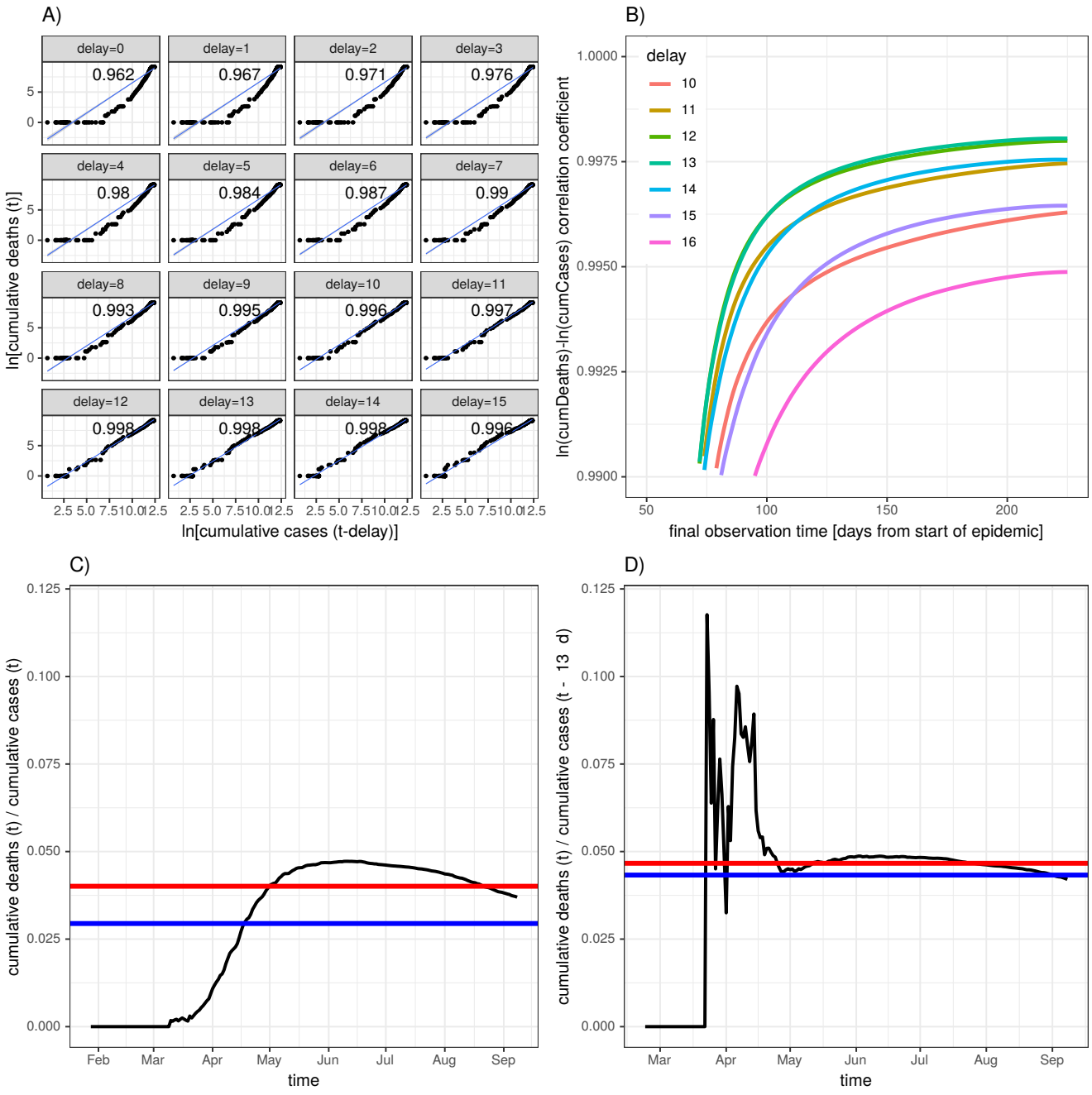

Figure 2. Asymptotic fatality-case ratio for the German Covid-19 data: A) Logarithmised cumulative deaths, $\ln(cumDeaths(t))$, versus time delayed logarithmised cumulative cases, $\ln(cumCases(t - delay))$, for different delays as indicated in the panel headers along with linear correlation (regression line plus Pearson's correlation coefficient printed in the upper left corner of each panel). B) Correlation coefficient as a function of the lengths of the time series (i.e., final observation time) for delays ranging from 10 to 16. C) Delay-0 asymptotic fatality-case ratio (black) with time average (blue) and median (red). D) Delay-13 asymptotic fatality-case ratio (black curve) with time average (0.043, blue line) and median (0.047, red line).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

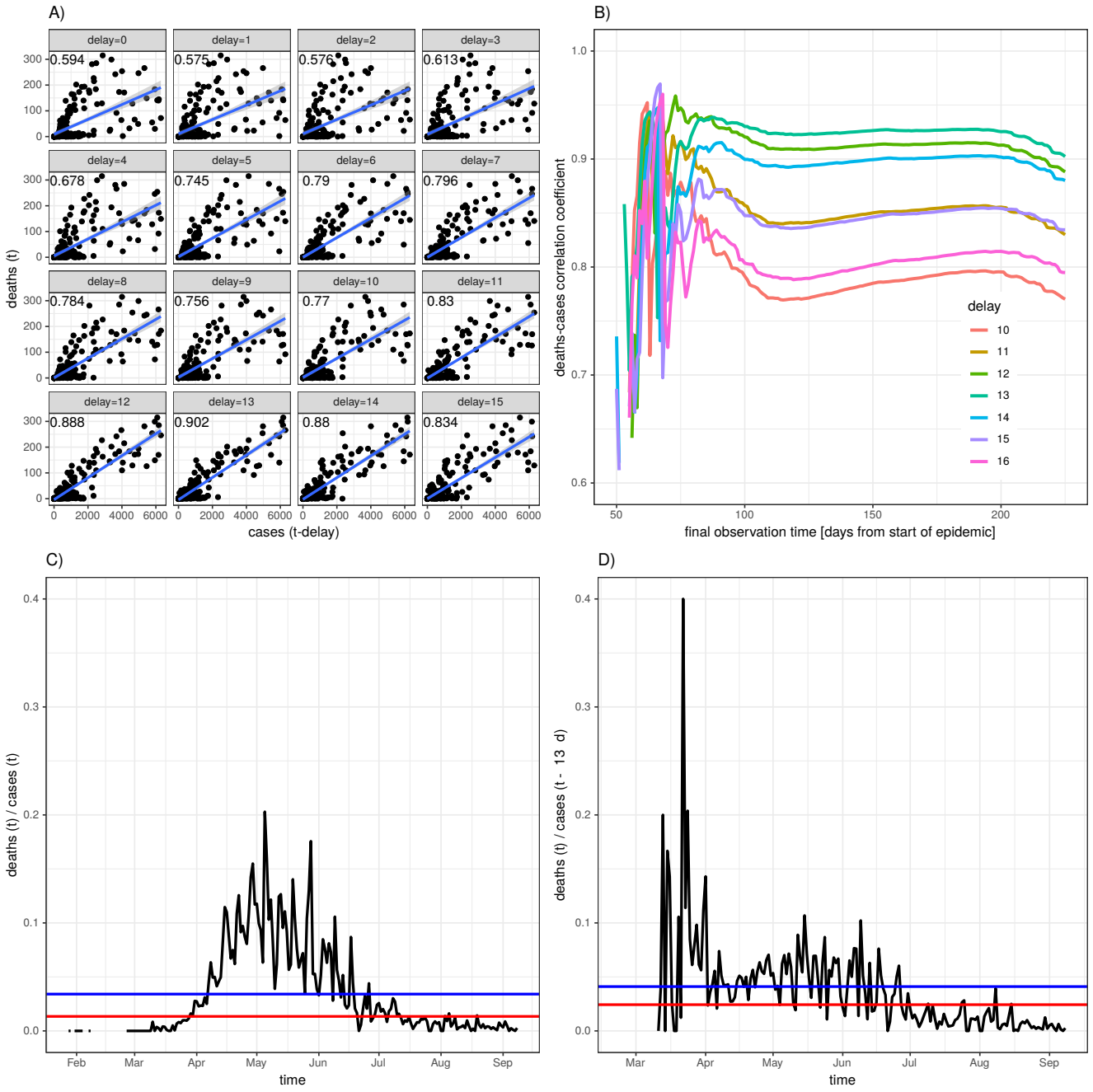

Figure 3. Instantaneous fatality-case ratio for the German Covid-19 data: A) New deaths ($deaths(t)$) versus time delayed new cases ($cases(t-delay)$) along with linear correlation (regression line plus Pearson's correlation coefficient printed in the upper right corner of each panel). B) Correlation coefficient as a function of the lengths of the time series (i.e., final observation time) for delays ranging from 10 to 16. C) Delay-0 instantaneous fatality-case ratio (black) with time average (blue) and median (red). D) Delay-13 instantaneous fatality-case ratio (black) with time average (0.041, blue) and median (0.024, red).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

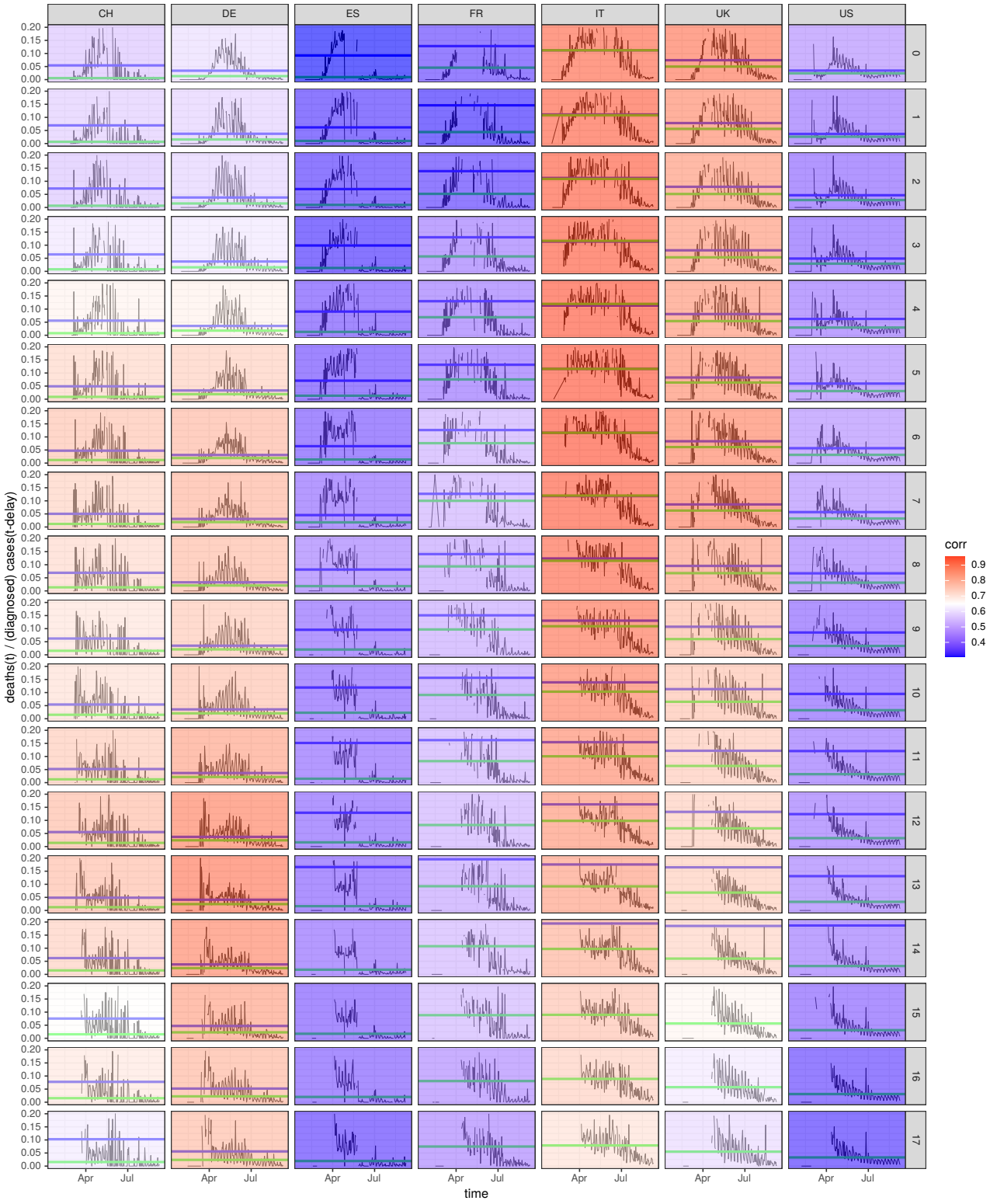

Figure 4. Diagnosis-to-death duration for 7 selected countries analysed using delay-time correlation: The plot shows the magnitudes of delay-specific correlations between $deaths(t)$ and $cases(t-\Delta t)$ for 7 selected countries (column labels) in form of a heatmap. The delays Δt (row labels) run from $\Delta t = 0d$ through $\Delta t = 17d$. Strong correlations are shown in dark red and declining correlation coefficients gradually fade to blue. Also shown for each country and each delay are the time courses of Delay$-\Delta t$ instantaneous fatality case ratios along with time average (blue line) and median (green).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

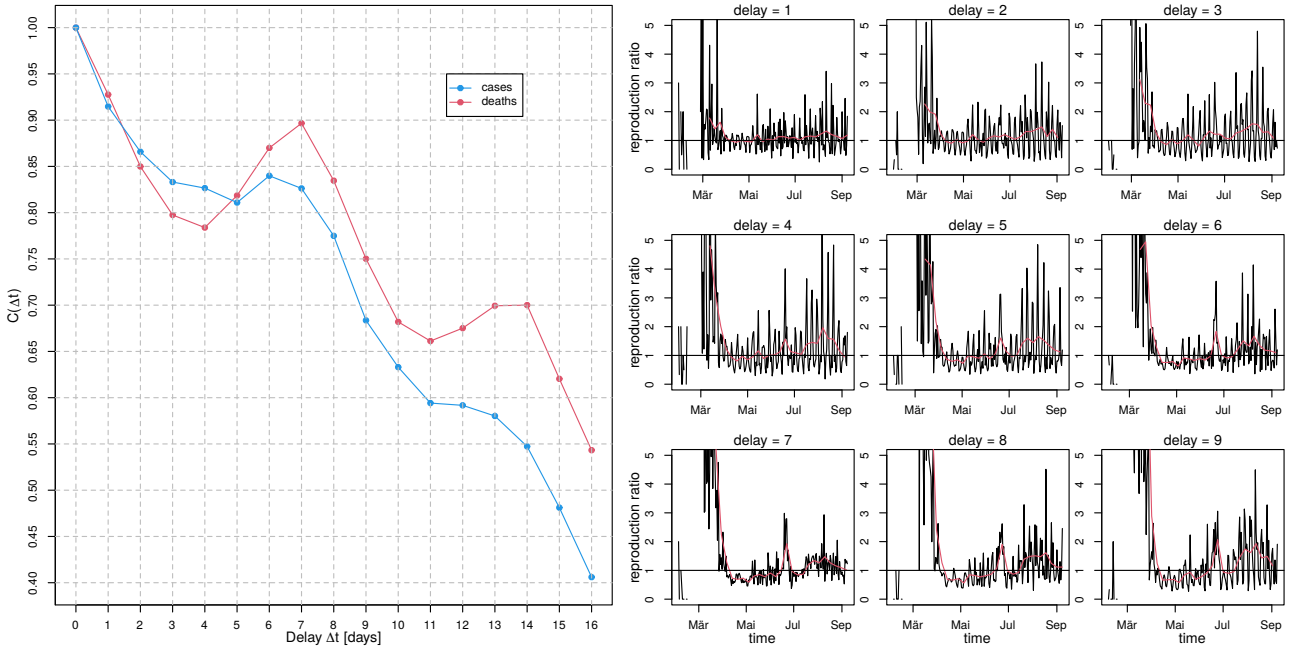

Figure 5. Delay-time autocorrelation for German incidence data: Left panel) Autocorrelation, $C(\Delta t)$, of cases (blue curve) and deaths (red) as a function of delay Δt. Right panel) Ratio $\frac{cases(t)}{cases(t-\Delta t)}$ (primitive approach to estimate the reproduction ratio) for 9 different delays Δt as indicated in the panel headers. The red curves result from a moving average with a window width of 7 days.

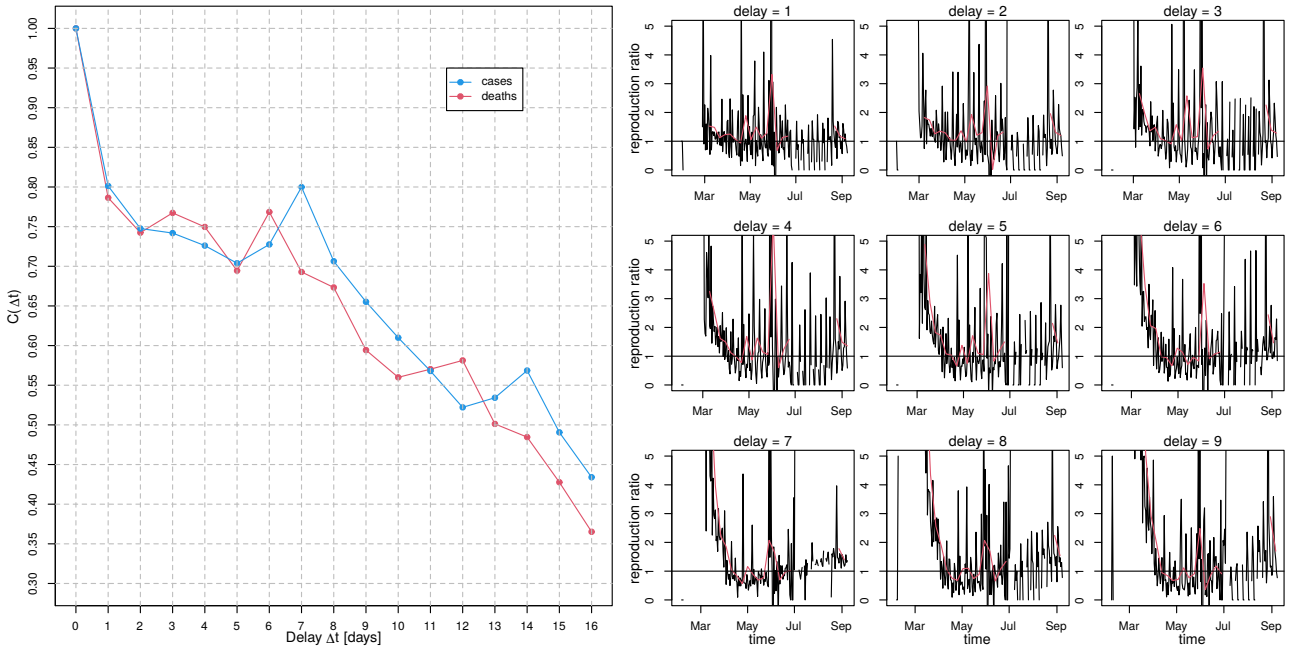

Figure 6. Delay-time autocorrelation for French incidence data: Left panel) Autocorrelation, $C(\Delta t)$, of cases (blue curve) and deaths (red) as a function of delay Δt. Right panel) Ratio $\frac{cases(t)}{cases(t-\Delta t)}$ (primitive approach to estimate the reproduction ratio) for 9 different delays Δt as indicated in the panel headers. The red curves result from a moving average with a window width of 7 days.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

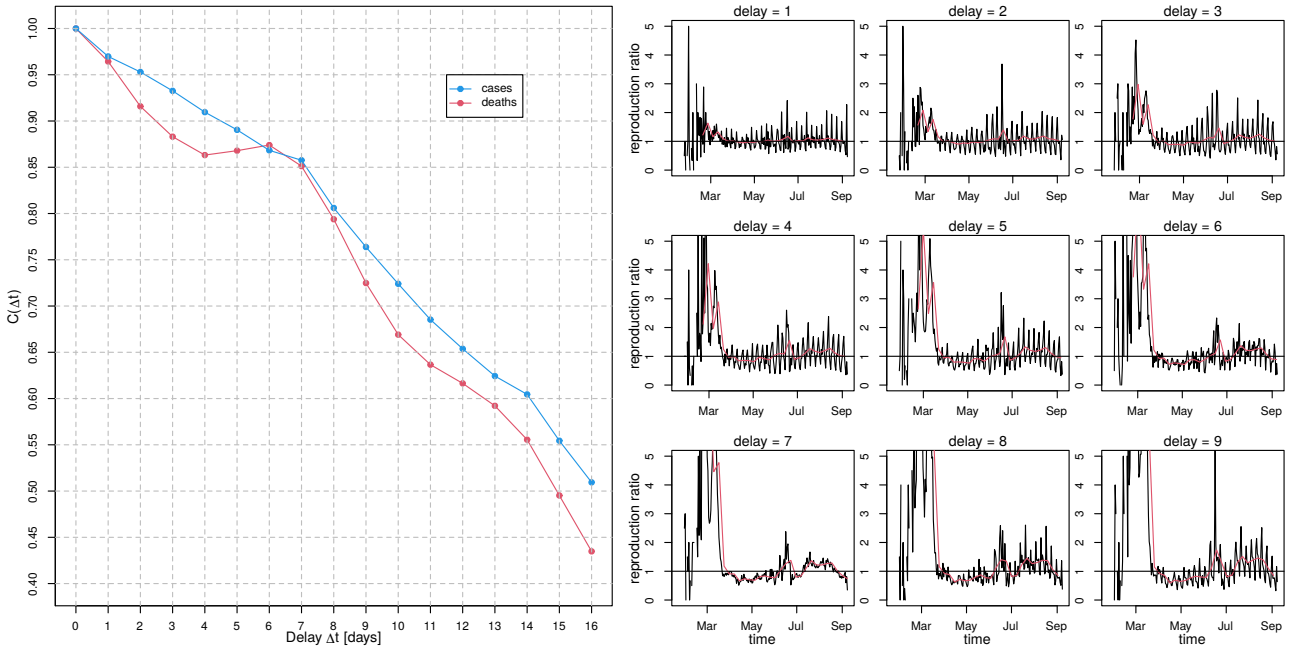

Figure 7. Delay-time autocorrelation for German incidence data based on the dataset provided by the RKI: This plot serves as comparison to fig. 5 to highlight differences between the ECDC and RKI datasets. Left panel) Autocorrelation, $C(\Delta t)$, of cases (blue curve) and deaths (red) as a function of delay Δt. Right panel) Ratio $\frac{cases(t)}{cases(t-\Delta t)}$ (primitive approach to estimate the reproduction ratio) for 9 different delays Δt as indicated in the panel headers. The red curves result from a moving average with a window width of 7 days.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

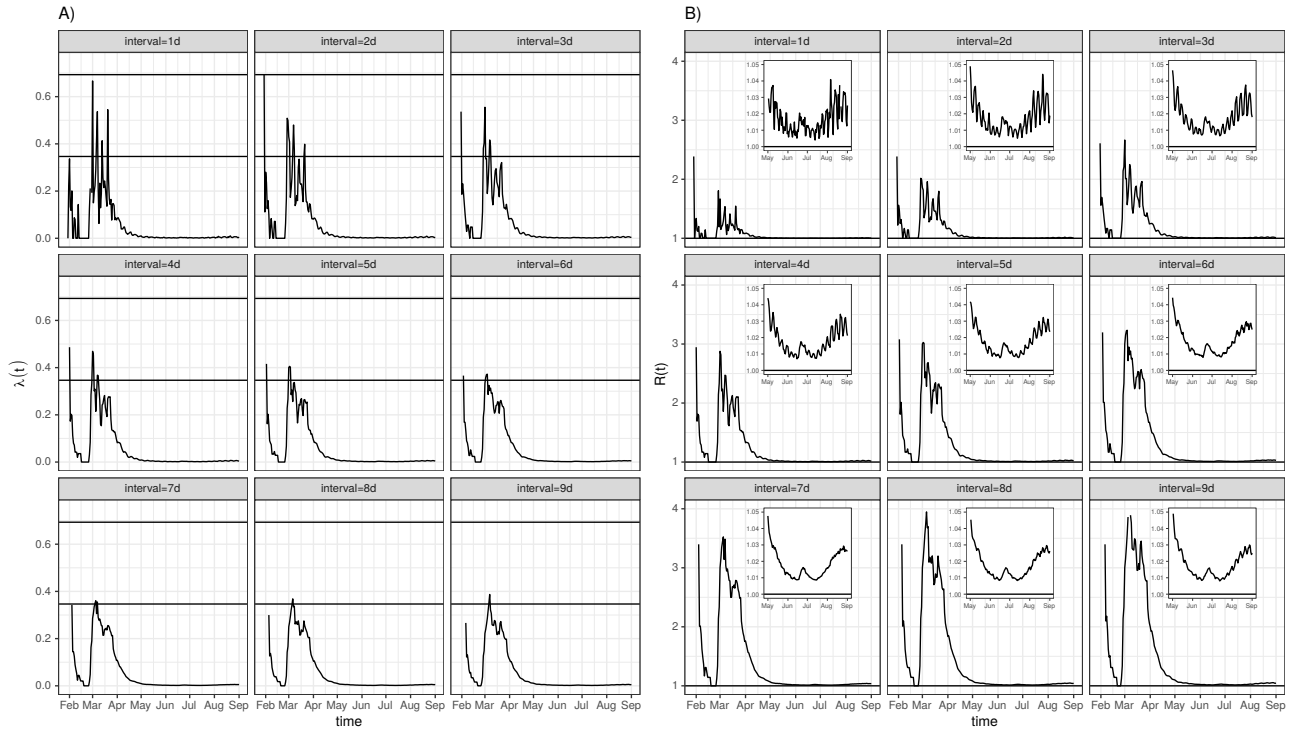

Figure 8. Time-dependent infection rate and approximate effective reproduction number for Germany: A) Time course of the infection rate $\lambda_{\Delta t}(t)$ according to eq. 7 for 9 different intervals (delays) Δt as indicated in the panel headers. Also shown are lines which correspond to doubling times of either $1d$ or $2d$, respectively. B) Approximate reproduction numbers calculated according to eq. 8. The inlets show details where R is close to 1, i.e. from May on. Of note, computed this way, R has a lower limit of 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2020 doi:10.20944/preprints202009.0628.v1

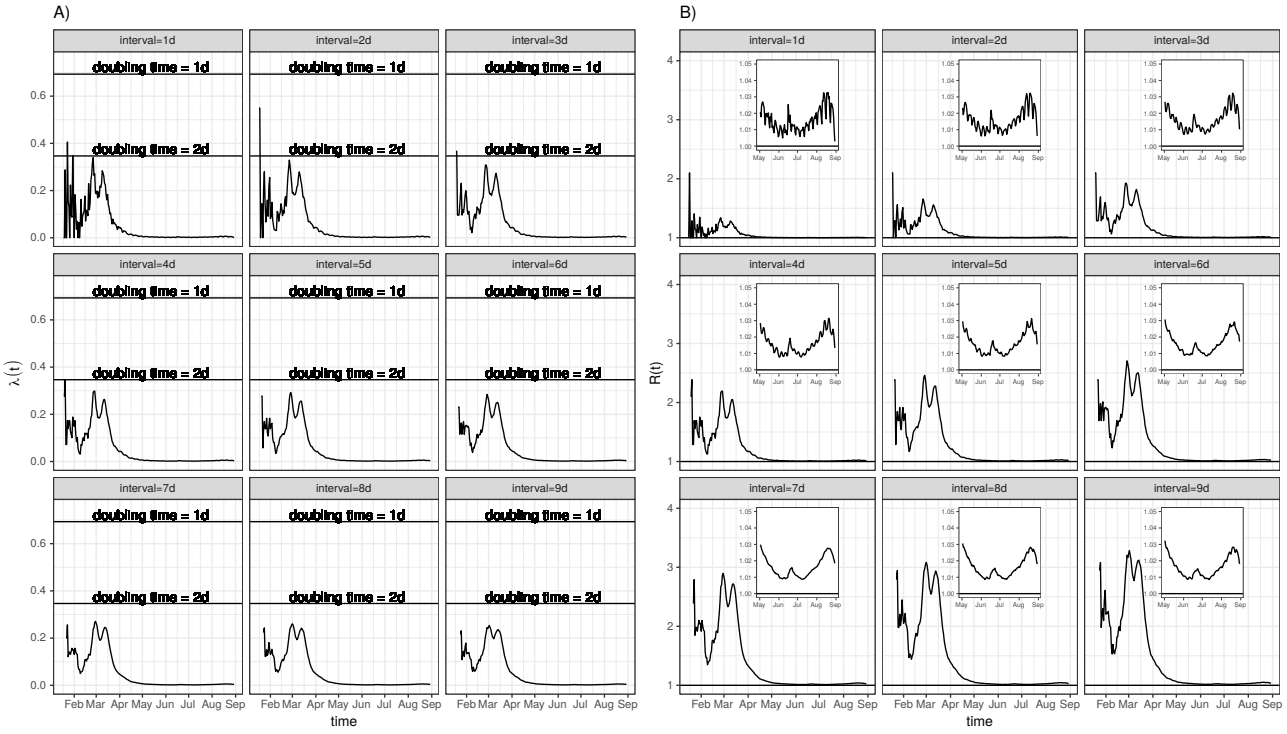

Figure 9. Time-dependent infection rate and approximate effective reproduction number for Germany based on the dataset provided by the RKI: This plot serves as comparison to fig. 8 to highlight differences between the ECDC and RKI datasets. A) Time course of the infection rate $\lambda_{\Delta t}(t)$ according to eq. 7 for 9 different intervals (delays) Δt as indicated in the panel headers. Also shown are lines which correspond to doubling times of either $1d$ or $2d$, respectively. B) Approximate reproduction numbers calculated according to eq. 8. The inlets show details where R is close to 1, i.e. from May on. Of note, computed this way, R has a lower limit of 1.