

## RESEARCH ARTICLE

# A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020

Marc Schneble<sup>1</sup>  | Giacomo De Nicola<sup>1</sup> | Göran Kauermann<sup>1</sup> | Ursula Berger<sup>2</sup>

<sup>1</sup> Department of Statistics, LMU Munich, Munich, Germany

<sup>2</sup> Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany

**Correspondence**

Marc Schneble, Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany.

Email:

[marc.schneble@stat.uni-muenchen.de](mailto:marc.schneble@stat.uni-muenchen.de)



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

**Abstract**

The case detection ratio of coronavirus disease 2019 (COVID-19) infections varies over time due to changing testing capacities, different testing strategies, and the evolving underlying number of infections itself. This note shows a way of quantifying these dynamics by jointly modeling the reported number of detected COVID-19 infections with nonfatal and fatal outcomes. The proposed methodology also allows to explore the temporal development of the actual number of infections, both detected and undetected, thereby shedding light on the infection dynamics. We exemplify our approach by analyzing German data from 2020, making only use of data available since the beginning of the pandemic. Our modeling approach can be used to quantify the effect of different testing strategies, visualize the dynamics in the case detection ratio over time, and obtain information about the underlying true infection numbers, thus enabling us to get a clearer picture of the course of the COVID-19 pandemic in 2020.

**KEYWORDS**

case detection ratio, COVID-19, dark figure of infections, generalized additive models, penalized splines

## 1 | INTRODUCTION

Originating from Wuhan, China, coronavirus disease 2019 (COVID-19) developed to become a worldwide pandemic in the spring of 2020 (Velavan & Meyer, 2020). Starting from the very beginning of this unprecedented health crisis, the issue of case detection, while always being at the center of scientific and public discourse, has been all but transparent. Knowing how many infections are really present in the population would be of paramount importance, and researchers have tried to tackle the problem in several different ways. Early in the epidemic wave, the ratio of undetected COVID-19 cases was likely to be high, that is, 5–20 times higher than the number of confirmed cases (e.g., Li et al., 2020 or Wu et al., 2020). The problem of discovering the case detection ratio (CDR) is tightly intertwined with the issue of uncovering the true fatality ratio of the disease, as knowledge on one of those two unknown quantities would provide information about the other. A natural experiment that allowed to obtain initial estimates of both the fatality ratio and the CDR occurred with the outbreak on the cruise ship “Diamond Princess” (Mizumoto et al., 2020). During the early stages of the pandemic, the actual percentage of the population infected for 11 European countries was deduced from early estimates of the mortality

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2020 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

rates (Flaxman et al., 2020). Moreover, Aspelund et al. (2020) used Bayes arguments applied to testing data from Ireland to estimate the CDR in the order of 7–11% at the beginning of the pandemic, and in the order of 10–20% after that. The argument is based on relating the number of tests and the share of positive tests. A similar approach has been pursued making use of Canadian data (Benatia et al., 2020). The problem of estimating the true numbers of COVID-19 infections has also been discussed from a purely statistical point of view, where the CDR was related to the fatality ratio (Manski & Molinari, 2020). A capture–recapture approach to estimate the total number of COVID-19 cases was proposed by Böhning et al. (2020) and Rocchetti et al. (2020), where the latter derive an upper bound for the cumulative number in mid-April for 10 European countries. The ratio of the upper bound and the observed number of cases ranges from around 4 (Greece) to around 8 (France). The capture–recapture method makes only use of publicly available data on COVID-19 cases and deaths, which also holds for the method that we present in this note. Here, we assume that the number of infected can be split into detected and undetected infections. In SIDARTHE models (Giordano et al., 2020), there is additional distinction into either asymptomatic or symptomatic cases, which we ignore here since the database that we use does not reliably contain these numbers. However, it should be noted that pre- and asymptomatic individuals have a significant impact on the spread of a pandemic disease, especially in the younger population (Stella et al., 2020). Thereby, presymptomatic individuals play a more significant role than asymptomatic ones (Buitrago-Garcia et al., 2020). Nonetheless, the number of asymptomatic cases can reduce the reproduction value of a disease because a background immunity is established, as shown for influenza transmission (Mathews et al., 2007).

Overall, underreporting appears to be an overarching problem, which plays a central role when estimating the CDR for COVID-19 (Russell et al., 2020). The importance of assessing the detection ratio and its effect on predictions of future infections has been demonstrated in mathematical simulation studies (Fuhrmann & Barbarossa, 2020). In this context, different national underreporting ratios have been compared (e.g., Rahmandad et al., 2020 or Jagodnik et al., 2020) and a general discussion and survey on assessing the infection fatality ratio (IFR) was conducted (Levin et al., 2020). In general, it is clear that the CDR changes greatly over time depending on testing strategy and capacities, which vary over time and across different regions. In Germany, the number of tests has increased considerably since the pandemic outbreak in March 2020. The testing strategy has also been adjusted several times: In the beginning, mainly individuals with symptoms were being tested, whereas in later phases, a very high number of tests have been performed on travelers returning from foreign countries and contact persons of COVID-19-positive individuals.

In this note, we explore the dynamics in the CDR using publicly available registry data on COVID-19 infections in Germany from March to December 2020 provided by the Robert-Koch-Institute (RKI). It is important to mention that in Germany's first months of the pandemic, no mass or systematic testing of the population had taken place. Our model therefore only makes use of a limited amount of information. We propose to jointly model fatal and nonfatal infections using a dynamic generalized linear mixed model with smooth random effects (see, e.g., Durbán et al., 2005; Durban & Aguilera-Morillo, 2017; Wood, 2017). The major advantage of our approach is that it only relies on the assumption that age-specific COVID-19 fatality ratios, while unknown, have not substantially changed over time. Whether this assumption is valid is currently discussed (Harris, 2020; Kip et al., 2020) and the possibility of differing fatality ratios in the second wave has been considered as well (Aspelund et al., 2020; Kenyon, 2020). To assess the impact of this assumption on our results, we provide sensitivity analyses and a simulation study in the Supporting Information, which demonstrate that our approach is sufficiently robust if there is no abrupt change in the infection fatality ratio.

Overall, our approach allows investigating the following. First, we explore how the case detection rate has changed over time, how it varies among different age groups, and if and how it changes in different regions of Germany, depending on infection dynamics and different testing strategies. Second, the model also provides an estimate of the dynamics in the true number of infections, regardless of whether they have been detected or not. All in all, this provides insight into the course of the COVID-19 pandemic, built exclusively on registry data.

The remainder of the paper is structured as follows. We describe the data constellation in depth in Section 2, and we propose our model in Section 3. In Section 4, we show the results of our analyses and provide extensive interpretations, whereas Section 5 concludes the paper with some implications and limitations of our study.

## 2 | DATA

We make use of COVID-19 data openly provided by the RKI, the German federal government agency and scientific institute responsible for health reporting, disease control, and prevention in humans (Esri Deutschland GmbH, 2020). The data, exemplified in Table 1, contain cumulated counts of newly registered, laboratory-confirmed COVID-19 cases in Germany

**TABLE 1** Illustration of the data structure. To facilitate reproducibility, the original column names used in the RKI dataset are given in brackets below our English notation

District (Landkreis)	Age group (Altersgruppe)	Gender (Geschlecht)	Cases (Anzahl Fall)	Deaths (Anzahl Todesfall)	Registration date (Meldedatum)
⋮	⋮	⋮	⋮	⋮	⋮
Munich City	60–79	F	26	0	September 8, 2020
Munich City	60–79	M	21	1	September 8, 2020
⋮	⋮	⋮	⋮	⋮	⋮

for each calendar day stratified by age group (0–4, 5–14, 15–34, 35–59, 60–79, or 80+ years), gender (male/female), and district (412 in total). Furthermore, for all registration dates and strata, the number of deaths associated with COVID-19 transmitted to the RKI by the local health authorities of the respective district is recorded. Note that the date of death is not provided, but for each death, we have the date when the infection was detected and confirmed by a (PCR) test. The database of the RKI is updated every morning with the new numbers transmitted to it from the local health authorities.

In this study, we only consider data entries with registration dates ranging from calendar week (CW) 10 (mid-March) to CW 53 (end of December) of the year 2020. For earlier weeks, the number of tests being positive was not large enough to draw conclusive results. On the other hand, the German vaccination campaign started at the very end of 2020. As this increasingly reduces the IFR, we only include infections that were registered in 2020. Consequently, the final outcome of almost all of these infections is known today. Moreover, although the data are given on a daily resolution, we here aggregate it into weekly data, which renders reporting delays occurring over the weekends and weekly reporting cycles irrelevant to our analysis, leading to more stable results. Since for children aged 14 years and younger, barely, any fatalities have been recorded, we excluded these age groups from our analysis.

To give a first insight into the data at hand, we plot in Figure 1 the raw numbers of cases reported by the official health authorities over time together with the raw number of fatalities stratified by age group. This is shown in the top four plots on a log-scale. Both the number of registered cases and that of fatal cases (indexed by registration date of the infection, and not by day of death) peak in CW 13 for the two younger age groups and in CW 14 for the two oldest age groups, respectively. Over the following weeks, these numbers decrease. The small peak in CW 25 was caused by an outbreak in the district of Gütersloh, which is explored in more depth later on in the paper. From CW 28 onward, we resume seeing an exponential increase of registered cases, whereas the numbers of registered fatal cases only start to rise 7 weeks later, also exponentially. By the end of the year 2020, we see a slight decrease in registered infections.

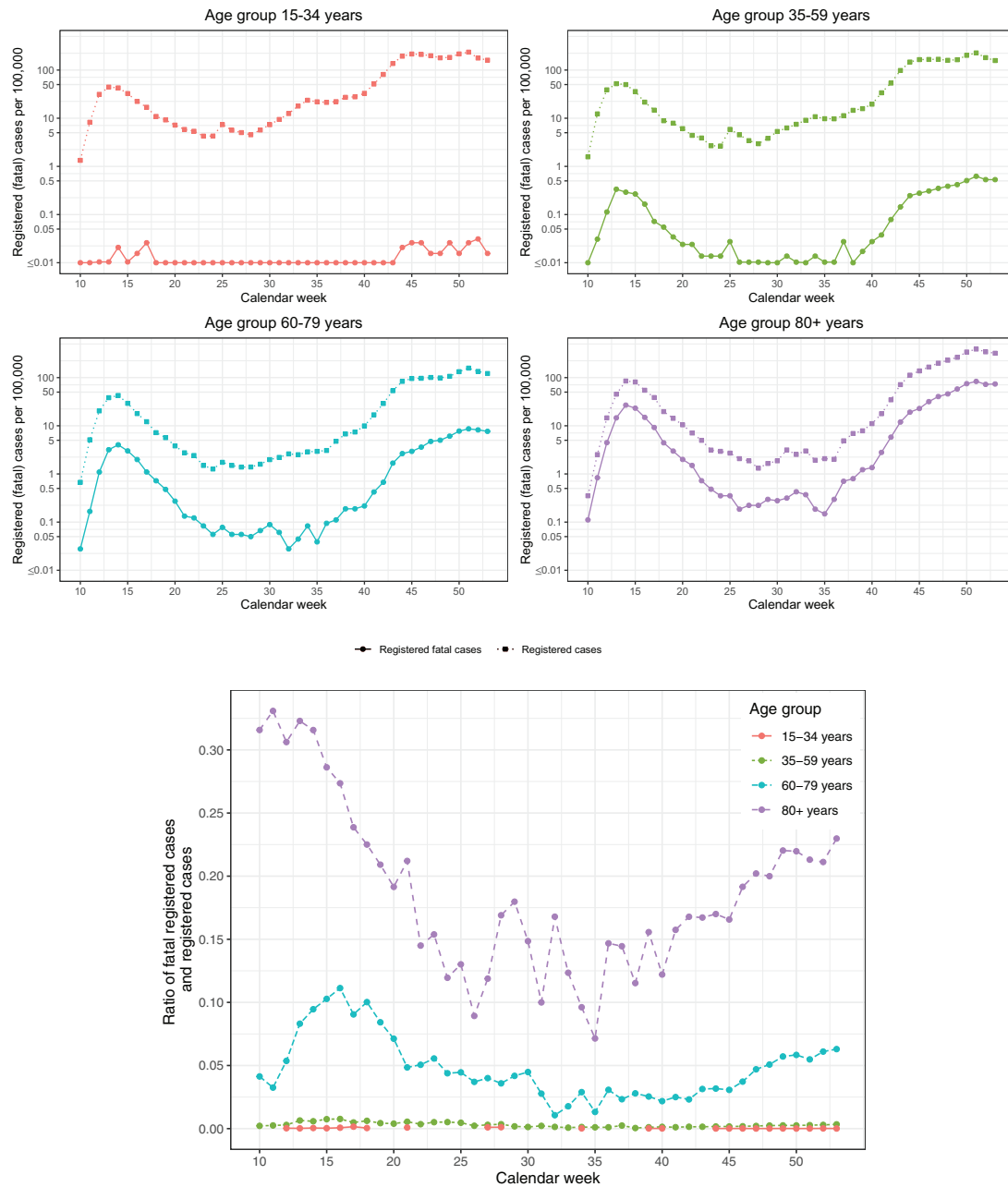
The raw case fatality ratio, calculated as the ratio of fatal cases over total registered cases, stratified by age group, is shown at the bottom of Figure 1. The raw case fatality ratio for the age group 80+ generally dropped from CW 10 onward and fluctuated mostly between 10% and 15% from week 25 onward. However, since CW 40 the case fatality ratio in this age group steadily climbed up to more than 20%. For the age group 60–79, the case fatality ratio has peaked in CW 16 and gradually decreased to 2.5%. Here, we also observe a steady increase toward the end of 2020, which results in more than a doubling of the case fatality ratio within 10 weeks. All other age groups exhibit relatively low raw case fatality ratios throughout.

Note that the raw data do not contain undetected cases, and therefore cannot provide a complete picture of the actual infection numbers, nor do these plots provide any information about the CDR. In the following, we develop a statistical model that enables us to estimate the relative changes in the CDR and the true infection numbers over time.

### 3 | METHODS

When describing the dynamics of the COVID-19 pandemic, the number of interest is the true count of newly infected persons in a cohort, which shall be denoted by  $I_t$  for week  $t = 1, \dots, T$ . Note that  $I_t$  remains unobservable. However, the number can be decomposed into the number of detected and reported cases  $D_t$  and the unknown number of newly infected persons, who have not been tested and remain undetected, which we can call the “dark number,”  $U_t$ . Hence, we have  $I_t = D_t + U_t$ , and  $D_t/I_t$  defines the CDR, which, however, remains unknown due to  $U_t$  being unknown.

Note that the index  $t$  indicates the time point on which the infection took place, which is usually unknown. The infection is eventually detected through a positive test at a later time point  $\tilde{t} = t + d$ . As  $d$  is often unknown, in particular, if the spread of the disease is diffuse, we will conceptually omit  $d$  in the following, which means that we set  $t$  equal to the



**FIGURE 1** Raw data: registered cases of COVID-19 infections and registered fatal cases on a weekly basis for Germany. Top figure: Absolute numbers on a log-scale stratified by age group. Bottom figure: Case fatality ratios (= fatal cases / registered cases) stratified by age group

registration date when an infection is confirmed through a test. This time point is the registration date described in the previous section. Generally, this approach is justifiable for COVID-19 infections because the range of delay  $d$  is small compared to the time range  $T$  of our data analysis (Mallett et al., 2020).

From today's perspective, we have uncensored knowledge on the outcomes of all reported cases  $D_t$ . That is, we know if they ended fatally or if they recovered. Consequently, the reported cases are composed of recovered (nonfatal) outcomes  $R_t$  and fatal outcomes  $F_t$ , that is,  $D_t = R_t + F_t$ . Given this, the total number of infected persons splits into  $I_t = R_t + F_t + U_t$ .

The expected number of reported fatal cases  $F_t$  as well as the expected number of recovered cases  $R_t$  are fractions of the total number of infections  $I_t$ . This leads to

$$\mathbb{E}(F_t | I_t) = I_t a \text{ and } \mathbb{E}(R_t | I_t) = I_t c_t, \quad (1)$$

where  $0 < (a + c_t) < 1$ . Here, quantity  $a$  defines the infection fatality ratio (IFR), whereas  $c_t$  is the CDR of nonfatal (recovered) infections. Note that these nonfatal infections also include mild and symptom-free cases. Thus, if testing capacities are increased or the testing strategy is changed,  $c_t$  will change as well, which is incorporated in the notation by time index  $t$ . In contrast, the IFR  $a$  will be assumed to remain constant over time. This can be justified by the fact that fatal cases, due to their severeness, are likely to be detected independently of any testing policy. This also includes, to some extent, postmortem tests.

With this notation, we obtain the time-dependent case detection ratio  $\text{CDR}_t = a + c_t$ . Note that for the dark number, that is, the latent number of undetected infections  $U_t$ , it holds that  $\mathbb{E}(U_t | I_t) = (1 - \text{CDR}_t)I_t$ . It would, of course, be favorable to estimate the number of undetected infections  $U_t$  via estimation of  $a$  and  $c_t$ . However, when only the reported fatal and nonfatal cases  $F_t$  and  $R_t$  are known, these two ratios cannot be estimated due to nonidentifiability issues, which we will demonstrate below. Nonetheless, with the data at hand, we are able to estimate the ratio  $c_t/a$ . To see this, we rewrite the above model in an equivalent form by defining a binary covariate  $x \in \{0, 1\}$  and by specifying the response variable  $Y_t$  through

$$Y_t | x = \begin{cases} F_t & \text{for } x = 0 \\ R_t & \text{for } x = 1. \end{cases}$$

This notational trick allows us to rewrite the above relations (1) as a regression model

$$\mathbb{E}(Y_t | I_t, x = 0) = \mathbb{E}(F_t | I_t) = \exp\{\log(I_t a)\} = \exp\{V_t + \alpha\}, \quad (2)$$

$$\mathbb{E}(Y_t | I_t, x = 1) = \mathbb{E}(R_t | I_t) = \exp\{V_t + \gamma_t\}, \quad (3)$$

where  $V_t = \log(I_t)$ ,  $\alpha = \log(a)$ , and  $\gamma_t = \log(c_t)$ . Equations (2) and (3) can, in turn, be summarized into a single regression model formula

$$\mathbb{E}(Y_t | V_t, x) = \exp\{V_t + \alpha + x(\gamma_t - \alpha)\}. \quad (4)$$

Note that  $I_t$  and hence  $V_t = \log(I_t)$  remain unobserved. We employ a Bayesian view and model  $V_t$  as normally distributed random effects  $V_t \sim N(\mu_t, \sigma^2)$ . Still, the parameters in model (4) are not identifiable, because any shift in  $\mu_t$  and a matching negative shift in  $\alpha$  and  $\gamma_t$ , respectively, results in the same model. This demonstrates the identifiability problem, which we have mentioned above. Hence, we are neither able to estimate the fatality ratio  $a = \exp(\alpha)$  nor the time-dependent ratio  $c_t = \exp(\gamma_t)$  with the data at hand. However, we can shift  $\mu_t$  such that the integral of  $\tilde{\mu}_t = \mu_t - k$  is equal to zero and define the global intercept  $\beta_0 = \alpha + k$ , which allows to rewrite (4) in an identifiable form (see Wood, 2017) to obtain the final regression model

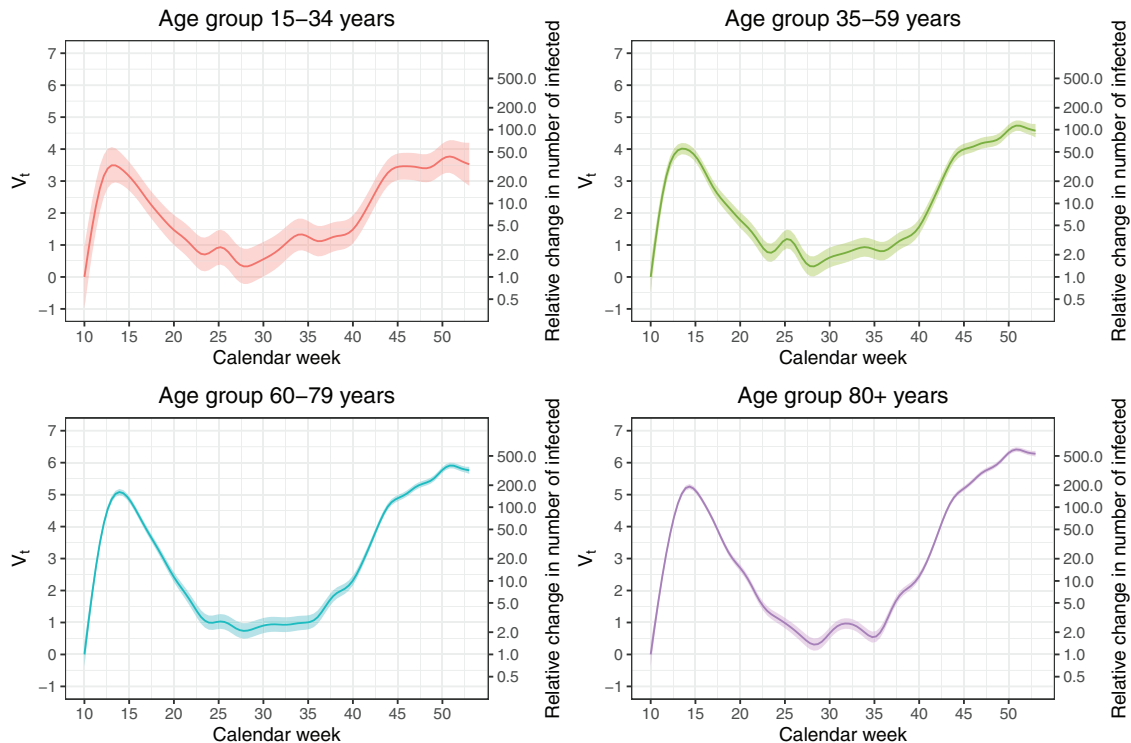
$$\mathbb{E}(Y_t | V_t, x) = \exp(V_t + \beta_0 + x\beta_t) \text{ and } V_t \sim N(\tilde{\mu}_t, \sigma^2) \text{ for } t = 1, \dots, T, \quad (5)$$

where  $\beta_t = \gamma_t - \alpha$  and  $\exp(\beta_t) = c_t/a$ . With this model, we can now explore the dynamics in the CDR. For two different time points  $t_1$  and  $t_2$ , we have using the small  $o()$  notation

$$\frac{\text{CDR}_{t_2}}{\text{CDR}_{t_1}} = \frac{c_{t_2} + a}{c_{t_1} + a} = \frac{c_{t_2}}{c_{t_1}} \{1 + o(a)\} = \frac{\exp(\beta_{t_2})}{\exp(\beta_{t_1})} \{1 + o(a)\} \approx \frac{\exp(\beta_{t_2})}{\exp(\beta_{t_1})}. \quad (6)$$

The latter approximation in (6) holds as long as the fatality rate  $a$  is small, which holds for COVID-19. Consequently,  $\beta_{t_2} - \beta_{t_1}$  can serve as a proxy for  $\log(\text{CDR}_{t_2}) - \log(\text{CDR}_{t_1})$ , and  $\exp(\beta_{t_2} - \beta_{t_1})$  is a proxy for the relative change in the case detection ratio  $\text{CDR}_{t_2}/\text{CDR}_{t_1}$ .

Based on these considerations, we see that it is necessary to model the dynamics in time  $t$  more appropriately to derive stable estimates for the CDR. It is natural to assume that changes in the CDR over time do not occur suddenly but gradually. For instance, test capacities are slowly increased and test strategies are gradually changed. To accommodate this in our model (5), we fit  $\beta_t$  by a smooth function in time leading to a time-varying coefficient model (Hastie & Tibshirani, 1993). We also induce smooth dynamics on the random component, leading to a time-varying random effect (Durban &



**FIGURE 2** Dynamics of the true infection numbers on the log-scale for different age groups: The smooth random effects  $V_t$ . The shaded areas represent 95% confidence bands

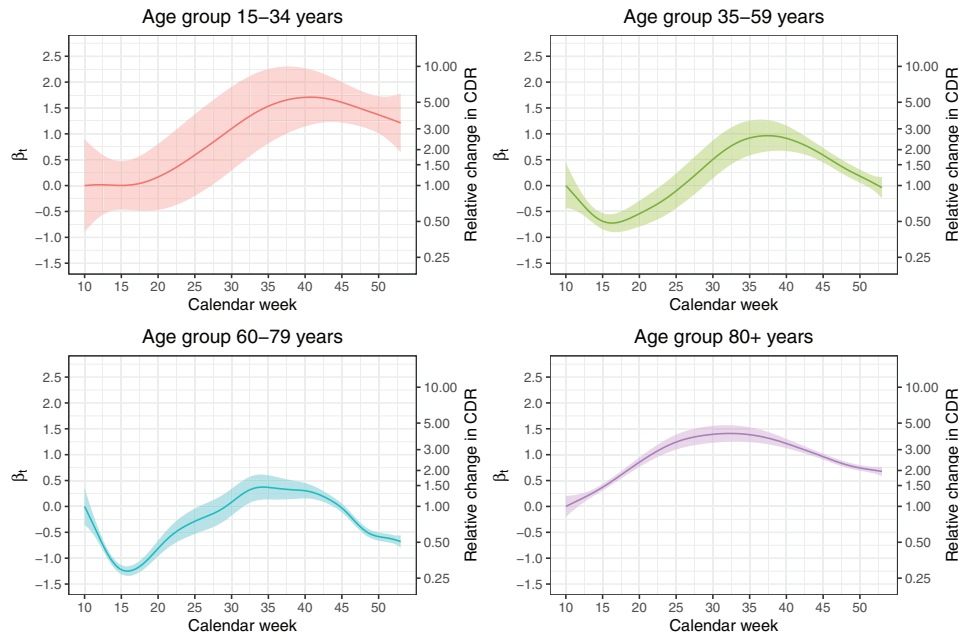
Aguilera-Morillo, 2017). These modifications lead to an identifiable and dynamic mixed regression model, for which we use a negative-binomial distribution for  $Y_t$  with a constant dispersion factor. The entire model can be fitted with standard software: All of our analyses were performed in **R** (R Core Team, 2013) and the dynamic mixed regression model is fitted using the **R**-package **mgcv** (Wood, 2017).

We apply this modeling approach using the reported data from CW 10 (beginning of March) up to CW 53 (final week of 2020), stratified by different age groups, to visualize the dynamics in the real infection numbers and the CDR from the beginning of the pandemic up to the beginning of the second wave. To assess the robustness of the approach concerning the assumption of time-constant and age-specific fatality ratios, we also refit the model when subdividing the data into different time frames. The results of this analysis are shown in the Supporting Information.

## 4 | RESULTS

### 4.1 | Model estimates

As the IFR  $a$  depends on age, we fit separate models for each of the relevant age groups defined by the RKI, that is, 15–34, 35–59, 60–79, and 80+ years. The dynamics in the true infection numbers on the log-scale, represented by the fitted smooth dynamic random effects  $V_t$ , are displayed in Figure 2. These curves mirror the relative change in the actual number of infected (detected and undetected) over time. Note that the absolute numbers cannot be interpreted on their own due to the mentioned identifiability issues. We therefore shift the curves such that  $V_{CW10} = 0$ . We can see that the relative course of the pandemic was very similar across all age groups, where a peak is reached around CW 14. However, the peak for the younger age groups is estimated to be around 1 week earlier than for the older age groups, that is, in CW 13. An explanation for this finding is that the younger age groups have been more affected by the lockdown, which started in Germany in CW 12. Looking at the difference between the maximum  $\max_t V_t$  and the minimum of  $V_t$  during the summer months, that is,  $\min_{20 \leq t \leq 40} V_t$ , we see that this difference increases with age, that is, the relative decline in true infections numbers after the first wave and the relative increase toward the second wave, respectively, was less pronounced in the younger age groups. Also eye-catching is the increase in infections around CW 25 for people below 60 years of age. This is



**FIGURE 3** Dynamics in the case-detection ratio for different age groups: The normalized time-varying coefficients  $\beta_t$ . The function values on the exp-scale (right y-axes) are the relative change in the case-detection ratio (CDR) with respect to calendar week 10

the aforementioned outbreak in the district of Gütersloh, which occurred in an industrial slaughterhouse and has mainly affected people of the working age. From CW 35 (end of August), all curves start rising steadily, where the steepest rise is seen for the oldest age group, whereas the rise is flatter for the younger age group. This shows that the second wave of the pandemic had already begun around CW 35. Moreover, Figure 2 shows that in all age groups but the youngest one, the peak of the second wave has surpassed the peak of the first wave.

Next, we look at the dynamics in the CDR. Figure 3 shows the fitted time-varying coefficients  $\beta_t$  together with corresponding 95% confidence bands. Again, the absolute level is not identifiable, so these curves are normalized such that  $\beta_{CW10} = 0$ . Hence, the function values on the exp-scale (right y-axes) give the relative change in the CDR with respect to CW 10. The CDR in the age group 80+ has risen monotonically since the beginning of the pandemic up to CW 33, where our model estimates the CDR to be more than four times higher as in mid-March. Note that in later weeks, the CDR among the elderly decreased again to the level of April/May. In contrast, for people aged 60–79, the CDR first dropped by about 70%, reaching its bottom as the pandemic passed its peak in Germany in CW 16. We subsequently see a monotonic increase, with the CDR becoming 1.5 times higher compared to the beginning of the pandemic. However, in this age group, the CDR has been more than halved from CW 40 up to the end of 2020 again. The dynamics in the CDR in the population aged 35–59 years are similar to those of the 60–79 years old: After a drop during March and April (CW 10–CW 16), the CDR increases, in mid-September, to nearly three times what it was in CW 10. For the youngest age group (aged 15–34), we also see a rise in the CDR over time, which seems substantial. However, the confidence bands in this age group are relatively wide because this age group is not as prone to fatal outcomes as older age groups.

## 4.2 | Interpretations

For the population aged 80 years and older, the CDR had increased until late summer, when it started to stagnate before slightly decreasing again. As the CDR can be at most 100%, and given that the relative change in this age group was about as high as a factor of 4 in CW 33 compared to March, we can conclude that at the beginning of the pandemic, the CDR among the population of 80 years and older could not have been more than 25%. Moreover, considering the relative change in the CDR, we can adjust the numbers from the peak in the first wave to be comparable, for example, to the numbers in week 40. To exemplify this, note that in week 40, the CDR for the age group 80+ was 2.3 times higher as in CW 15, at the peak of the first wave. This ratio results from the plot in Figure 3 (bottom right) by taking  $\beta_{CW15} = 0.4$  and  $\beta_{CW40} = 1.25$  and calculating the ratio  $\exp(1.25 - 0.4) = 2.3$ . In week 40, we had about 11 new infections per week per 100,000 reported

in this age group. In CW 15, this number had become 80. However, in week 15, the CDR was much lower as in CW 40, and thus, we would have seen  $2.3 \cdot 80 = 184$  cases per 100,000 in this age group 80+ if we had the same CDR in CW 15 as in CW 40.

For the population aged 60–79 years, the CDR between the minimum in CW 16 and its maximum in calendar week 34 changed by a factor of around 5. From this, we can deduce that around the peak of the first wave in Germany, at most 20% of the infections were detected, whereas at least 80% remained unseen. To be able to compare numbers from the first wave to those in autumn, we apply a similar calculation as above. This results in an estimated number of at least  $5 \cdot 17 = 85$  cases per 100,000, where only 16 cases per 100,000 have been observed in CW 16.

In the age group 35–59, the relative change of CDR during the minimum in CW 16 and the maximum in CW 36 was as high as a factor of 5 as well. Again, the same calculation shows that the 22 detected infections per 100,000 in week 16 would increase to  $5 \cdot 22 = 110$  cases per 100,000 if we would have had the CDR in week 16 as it was in week 36.

A general question in the pandemic is whether extensive testing leads to a high CDR. Applying our model to regional data allows us to investigate this question. The Supporting Information compares separate model fits for the two most populous German states, North-Rhine-Westphalia and Bavaria. The two states implemented different testing strategies over the summer months. Although in Bavaria, public test stations were opened in summer, particularly at the borders on the motorways, such fine screening of holiday returnees was not pursued in North-Rhine-Westphalia. Our model allows assessing and, in particular, quantifying how such different testing strategies lead to different CDRs in these two regions. The results quantify by how much the dark figure was reduced in relationship with the Bavarian testing strategy.

## 5 | DISCUSSION

Raw reported case numbers and measures derived from them, such as the case fatality ratio, are prone to changes in testing strategies and test capacities, which also influence the CDR. Comparisons between raw case numbers over time therefore need to be interpreted with care. The case-fatality ratio, calculated from the raw number of reported deaths related to COVID-19 divided by reported cases, is also impaired because deaths occur with a time delay after registration, meaning that deaths registered today correspond to infections that have been reported up to several weeks ago. Our method allows us to uncover relative changes in the CDR over different pandemic phases. Moreover, by shedding light on the number of undetected cases, we can describe the dynamics in the true number of COVID-19 infections for Germany from March 2020 until December 2020. The approach is based on publicly available data on registered cases and does not rely on simulations or additional survey data. We make use of the fact that, for each fatal outcome, the registration date of the infection is included in the data. This allows us to jointly model the number of registered nonfatal cases and that of fatal infections in a dynamic mixed model, leading to an assessment of the dynamics taking place in real infection numbers. Based on the available information on the relative change in the CDR over time, we are able to compare numbers from the first wave of the pandemic in spring with numbers from the second wave in autumn, adjusting for the difference in the proportion of undetected cases.

A general limitation of our approach is that it suffers from an identifiability issue and hence does not derive absolute values of the CDR. One may, however, combine our results with findings from seroepidemiological studies, which aim to assess the prevalence of COVID-19 in the general population by screening a representative sample. A list of current seroepidemiological studies in Germany is provided by the RKI (Robert-Koch-Institute, 2020). Although these studies provide crucial information on the current situation of the spread of the disease, they can only give a snapshot of the instantaneous situation when the study was conducted. With the knowledge of the dynamics in new infections given by our approach, the findings of such studies can be used to estimate the situation at other time points. For example, we look at the Prospective Covid-19 Cohort Study Munich (KoCo19, Radon et al., 2020). They report a CDR of about 25%, where the survey was run between May and June 2020 in the city of Munich. We can deduce that the CDR for October to be about three times higher for the 35–59 age group. More precise calculations would require age-specific numbers in the study as well as a regional refit of our model. A nationwide seroprevalence study was conducted between the beginning of July and mid-August of 2020, which yielded a CDR of around 55% in the adult population (ifo Institut & forsa, 2020). Nonetheless, the authors admit that the fading of COVID-19 antibodies could influence their findings sometime after the infection. A seroprevalence study, which is also nationwide but on a larger scale, is currently being carried out, but the results are not yet available.<sup>1</sup> In principle, however, this demonstrates that the combination of seroepidemiological studies and our approach allows obtaining estimates for absolute numbers of the CDR instead of relative comparisons only.

<sup>1</sup> [https://www.rki.de/DE/Content/Gesundheitsmonitoring/Studien/lid/lid\\_node.html;jsessionid=02C6FAB6F407B92315BDA5C1650F4D3A.internet072](https://www.rki.de/DE/Content/Gesundheitsmonitoring/Studien/lid/lid_node.html;jsessionid=02C6FAB6F407B92315BDA5C1650F4D3A.internet072)



A critical assumption of our model is that we assume the IFR  $a$  to be constant over time for a given age group and negligibly small compared to the detection ratio of nonfatal cases. The latter is certainly valid for the numbers we looked at. Staerk et al. (2021) show that most of the dynamics in the effective IFR of the German population can be explained by the varying age distribution of COVID-19 cases. As the age distribution within the RKI age categories varies as well, the IFR  $a$  within each age group might slightly change over time that, however, occurs not abruptly but smoothly over time. The sensitivity analysis, which can be found in the Supporting Information, provides evidence that our assumption of  $a$  being constant is, for the most part, fulfilled. With increasing vaccination levels in the population starting from January 2021, the assumption of a constant case fatality ratio becomes invalid. This eventually prevents the application of our model to later stages of the pandemic.


## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at <https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74>.

## OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

Marc Schneble  <https://orcid.org/0000-0001-9523-4173>

## REFERENCES

- Aspelund, K., Droste, M., Stock, J. H. & Walker, C. D. (2020). Identification and estimation of undetected COVID-19 cases using testing data from Iceland. NBER Working Paper w27528.
- Benatia, D., Godefroy, R., & Lewis, J. (2020). Estimates of COVID-19 cases across four Canadian provinces. *Canadian Public Policy*, 46(S3), S203–S216.
- Böhning, D., Rocchetti, I., Maruotti, A., & Holling, H. (2020). Estimating the undetected infections in the COVID-19 outbreak by harnessing capture–recapture methods. *International Journal of Infectious Diseases*, 97, 197–201.
- Buitrago-Garcia, D., Egli-Gany, D., Counotte, M. J., Hossmann, S., Imeri, H., Ipekci, A. M., Salanti, G., & Low, N. (2020). Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS Medicine*, 17(9), e1003346.
- De Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, 6(1), 50–62.
- Drikvandi, R., Verbeke, G., & Molenberghs, G. (2017). Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics*, 73(1), 63–71.
- Durban, M., & Aguilera-Morillo, M. C. (2017). On the estimation of functional random effects. *Statistical Modelling*, 17(1–2), 50–58.
- Durbán, M., Harezlak, J., Wand, M., & Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24(8), 1153–1167.
- Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2), 89–121.
- Esri Deutschland GmbH. (2020). Daily COVID-19 case numbers provided by the Robert-Koch-Institute. <https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74>
- Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Ghani, C. A., Donnelly, A. C., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C., & Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820), 257–261.
- Fuhrmann, J., & Barbarossa, M. V. (2020). The significance of case detection ratios for predictions on the outcome of an epidemic - A message from mathematical modelers. *Archives of Public Health*, 78(63).
- Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, 26(6), 855–860.
- Harris, J. E. (2020). COVID-19 case mortality rates continue to decline in Florida. *medRxiv*.
- Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4), 757–779.

- ifo Institut, & forsa. (2020). Die Deutschen und Corona - Schlussbericht der BMG-“Corona-BUND-Studie”. <https://www.ifo.de/publikationen/2020/monographie-autorenschaft/die-deutschen-und-corona>
- Jagodnik, K. M., Ray, F., Giorgi, F. M., & Lachmann, A. (2020). Correcting under-reported COVID-19 case numbers: Estimating the true scale of the pandemic. *medRxiv*.
- Kenyon, C. (2020). Flattening-the-curve associated with reduced COVID-19 case fatality rates-an ecological analysis of 65 countries. *Journal of Infection*, 81(1), e98–e99.
- Kip, K. E., Snyder, G., Yealy, D. M., Mellors, Minnier, T., Donahoe, M. P., McKibben, J., Collins, K., & Marroquin, O. C. (2020). Temporal changes in clinical practice with COVID-19 hospitalized patients: Potential explanations for better in-hospital outcomes. *medRxiv*.
- Levin, A., Hanage, W., Owusu-Boaitey, N., Cochran, B., Walsh, S. P., & Meyerowitz-Katz, G. (2020). Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *European Journal of Epidemiology*, 35, 1123–1138.
- Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490), 489–493.
- Mallett, S., Allen, A. J., Graziadio, S., Taylor, S. A., Sakai, N. S., Green, K., Suklan, J., Hyde, C., Shinkins, B., Zhelev, Z., Peters, J., Turner, P. J., Roberts, N. W., di Ruffano, L. F., Wolff, R., Whiting, P., Winter, A., Bhatnagar, G., Nicholson, B. D., & Halligan, S. (2020). At what times during infection is SARS-CoV-2 detectable and no longer detectable using rt-pcr-based tests? A systematic review of individual participant data. *BMC Medicine*, 18.
- Manski, C. F., & Molinari, F. (2020). Estimating the COVID-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics*, 220, 181–192.
- Mathews, J. D., McCaw, C. T., McVernon, J., McBryde, E. S., & McCaw, J. M. (2007). A biological model for influenza transmission: pandemic planning implications of asymptomatic infection and immunity. *PLoS One*, 2(11), e1220.
- Mizumoto, K., Kagaya, K., Zarebski, A., & Chowell, G. (2020). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance*, 25(10), 2000180.
- R Core Team. (2013). R: A language and environment for statistical computing.
- Radon, K., Saathoff, E., Pritsch, M., Guggenbühl, N., Jessica, M., Kroidl, I., Olbrich, L., Thiel, V., Diefenbach, M., Riess, F., Forster, F., Theis, F., Wieser, A., Hoelscher, M., Bakuli, A., Eckstein, J., Froeschl, G., Geisenberger, O., Geldmacher, C. . . . Schwetmann, L. (2020). Protocol of a population-based prospective COVID-19 cohort study Munich, Germany (KoCo19). *medRxiv*.
- Rahmandad, H., Lim, T. Y., & Sterman, J. (2020). Estimating COVID-19 under-reporting across 86 nations: Implications for projections and control. Available at SSRN 3635047.
- Robert-Koch-Institute. (2020). Seroepidemiological studies in the general population. [https://www.rki.de/EN/Content/infections/epidemiology/outbreaks/COVID-19/AK-Studien-english/Sero\\_General.html](https://www.rki.de/EN/Content/infections/epidemiology/outbreaks/COVID-19/AK-Studien-english/Sero_General.html)
- Rocchetti, I., Böhning, D., Holling, H., & Maruotti, A. (2020). Estimating the size of undetected cases of the COVID-19 outbreak in Europe: An upper bound estimator. *Epidemiologic Methods*, 9(s1).
- Russell, T. W., Hellewell, J., Abbott, S., Jarvis, C., van Zandvoort, K., Ratnayake, R., CMMID nCov working group, Flasche, S., Eggo, R., Edmunds, W. J., & Kucharski, A. J. (2020). *Using a delay-adjusted case fatality ratio to estimate under-reporting*. Centre for Mathematical Modeling of Infectious Diseases Repository.
- Staerk, C., Wistuba, T., & Mayr, A. (2021). Estimating effective infection fatality rates during the course of the COVID-19 pandemic in Germany. *BMC Public Health*, 21(1073).
- Stella, L., Martínez, A. P., Bauso, D., & Colaneri, P. (2020). *The role of asymptomatic individuals in the covid-19 pandemic via complex networks*. arXiv preprint arXiv:2009.03649.
- Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical Medicine & International Health*, 25(3), 278–280.
- Wood, S. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall/CRC Texts in Statistical Science. CRC Press.
- Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516), 1548–1563.
- Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Reingold, A., Arnold, B. F., Hubbard, A., & Benjamin-Chung, J. (2020). Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications*, 11(1), 1–10.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Schneble, M., De Nicola, G., Kauermann, G., & Berger, U. A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020. *Biometrical Journal*. 2021;63:1623–1632. <https://doi.org/10.1002/bimj.202100125>