Supplementary material

A Model details

The model defined by (5) fits into the comprehensive class of generalized additive mixed models (GAMM), which are generalized additive models (GAM) in which the linear predictor of the model also includes random components (Wood, 2017). In particular, the model assumes that

\[ Y_t | V_t, x \text{ indep.} \sim \text{NegBin}(\lambda_t, \phi), \]

where \( \lambda_t = E(Y_t | V_t, x) \), \( \phi \) is a constant overdispersion factor such that \( \text{Var}(Y_t) = \lambda_t + \phi \lambda_t^2 \), and \( V_t \) are random effects with \( V_t \text{ indep. } N(\mu_t, \sigma^2) \). The conditional mean is modeled in log-linear fashion according to

\[ \log(\lambda_t) = V_t + \beta_0 + x \beta_t. \]

Here, \( \beta_0 \) is an intercept and \( \beta_t \) is the linear effect of \( x \), which varies over time. In order to model this time-varying coefficient smoothly, we represent \( \beta_t \) through a basis of B-splines where each B-spline is only supported locally (De Boor, 1972). We choose \( k = 30 \) B-splines, which would usually result in very wiggly estimates for \( \beta_t \). However, adding a penalty for the differences of B-spline coefficients to the log-likelihood criterion of the model as proposed by Eilers et al. (1996) allows to control the smoothness of \( \beta_t \). Therefore, the according smoothing parameter needs to be chosen appropriately, with this choice being the central part of GAMM model fitting routines (see Wood et al., 2016 for a general framework for smoothing parameter estimation in GAMMs). Penalized splines can be included in the linear predictor of the model using the \( s() \) function of the \texttt{mgcv} package by setting the argument \( \text{bs} = "ps" \). The random effects \( V_t \) are actually defining a random smooth curve which can also be included in the model with the \( s() \) function and by specifying \( \text{bs} = "fs" \). Here, we also set the maximal degree of freedom to \( k = 30 \) which is, however, again controlled by a single smoothing parameter. The model formula is supplied to the \texttt{gam()} routine which fits the model and supplies the model results as shown in Figures 2 and 3.

B Empirical estimates

A simple empirical estimate of the ratio \( \exp(\beta_t) = c_t / a \) can be obtained by replacing the expected values in (1) with observations. This suggests the raw estimate

\[ \hat{\frac{c_t}{a}} = \frac{R_t}{F_t} / \frac{I_t}{F_t} = \frac{R_t}{F_t}. \]  

Consequently, a raw estimate of the ratio of the CDR at two time points \( t_1 < t_2 \) as in (6) is given through

\[ \hat{\frac{c_{t_2}}{c_{t_1}}} = \frac{R_{t_2}/F_{t_2}}{R_{t_1}/F_{t_1}}. \]

We plot the raw estimates (7) for all considered age groups in Figure 4, which shows that the weekly fluctuation is quite large. Moreover, these estimates are not indicative of actual information about \( U_t \).

C Region specific analyses

Figure 5 shows that the Bavarian testing strategy succeeded in reducing the dark figure for the intermediate age groups (35-59 and 60-79 years old) when compared to North-Rhine-Westfalia. For the age group 80+ years, the course of the case detection ratio is instead practically identical in these two federal states. However, this is not surprising since the additional tests aimed to detect infected returnees from their holidays and the age group 80+ is the minor mobile age group.
D Sensitivity analysis

The analysis that we have conducted above has shown that, in Germany, the first COVID-19 wave has peaked around calendar week 14. This became noticeable through higher infection numbers (Figure 2). At the same time, we can see an estimated local minimum in the case detection ratio for all age groups but the oldest one (Figure 3). To assess the robustness of our approach and to validate the crucial assumption that the IFR is constant for a given age group, we now repeat the analysis for three different subperiods. In particular, those are the first wave (CW 10 - CW 22), the plateau on a low level after the first wave (CW 23 - CW 35) and the time from the beginning of the second wave until the end of the year 2020 (CW 36 - CW 53). This division becomes apparent when considering Figure 1 or Figure 2, respectively. Note again that we fit a separate model for every age group. Figure 6 shows the previous Figure 3 overlaid by the fitted $\beta_t$ as dotted lines when only fitting the model with data from each of the three respective time windows. We see that for all age groups but the youngest one, the dotted curve follows the $\beta_t$, which results from the complete data model fit or even lies within the respective confidence bands. Note that gam often reduces the degrees of freedom of a smooth effect to one if the actual effect is close to linearity. This results in a fitted straight line which can be seen most prominently for the intermediate time frame. The data relating to the age group 15-34 is very sparse, i.e. very few deaths are observed, which results in wide confidence bands. When reducing the time window of the data as in our sensitivity analysis, this also introduces difficulties in the estimation of the smoothing parameter, which here results in either very wiggly estimates (the smoothing parameter chosen is too low) or in a fitted straight line (the smoothing parameter is too high). All things considered, the results seen in Figure 6 suggest that the assumption of a time-constant age-specific IFR is not violated for Germany.

Figure 4 Empirical estimates of the ratio $c_t/a$ for each age group.
Figure 5  Dynamics in the case-detection ratio for Bavaria (left panels) and North-Rhine-Westphalia (right panels): 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.

E  Simulation study

In order to validate the statistical model developed in Section 3, we conduct a simulation study for a single age group. First, we want to assess the ability of the model for estimating the unobserved log-infection numbers, for which we assumed the prior distribution $V_t \sim N(\mu_t, \sigma^2)$ before making the model identifiable. The IFR is here set to $a = 0.02$, the time varying case detection ratio of recovered cases is assumed to be $c_t = \cos(t/10 + 2 \cdot \pi/3) + 1.5)/4$ for $t = 1, \ldots, T = 30$ and the random effects distribution has variance $\sigma^2 = 0.3$. Furthermore, we set $E(V_t) = \mu_t = 2.5^{1.2} \exp(-0.15t) + \exp(1 + 0.09(t - 20))$, which mirrors a first wave as well as the beginning of a second wave. The following data-generating
Figure 6  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. Solid lines including 95% confidence bands: Model includes all data; dotted lines: Model includes data delimited by the vertical dashed lines.

process is repeated 1000 times. For $t = 1, \ldots, T$ we simulate $V_t \sim N(\mu_t, \sigma^2)$ such that the true infection numbers are given by $I_t = \exp(V_t)$. Thus, fatal and detected recovered infections are simulated according to $F_t \sim \text{Bin}(I_t, a)$ and $R_t \sim \text{Bin}(I_t, c_t)$, respectively. We fit the model for each of the $N$ scenarios and obtain estimates $\hat{V}_t$ for the centered $\tilde{\mu}_t$. Note again that these estimates are not identifiable, i.e. we can only identify relative changes on the exp-scale.

The left panel of Figure 7 shows the simulation results for the course of the true infection numbers on the log-scale. Here, the solid red curve shows $\mu_t - k$, where $k$ is the mean of the $\mu_t$, i.e. $\tilde{\mu}_t$ is centred around zero. The fitted values $\hat{V}_t$ are already centred around zero, as performed by the gam modelling routine, and the blue dashed curve shows the mean of these estimates over all simulations. The grey areas show the intervals for the respective quantiles 25%-75%, 10%-90% and 1%-99%. Overall, we can see that the model is able to capture the relative trend of the actual infection numbers very well.

Next, we want to assess the liability of the random effects distribution. We simulate $V_t$ from a mixture of three normal distributions such that the mean of $V_t$ equals $\tilde{\mu}_t$ and the variance is as least as large as above for all $t$. The result is shown in the right panel of Figure 7, which indicates that misspecification of the random effects distribution does only slightly affect the estimates of $V_t$ in terms of bias and variance. From a more general point of view, statistical tests for the validity of the assumption of normality of random effects are discussed in Drikvandi et al. (2017).
Figure 7  Results of $N = 1000$ simulations. The solid red curve shows the true mean of the random effects and the dotted blue curve shows the mean of the estimates $\hat{V}_t$ over all simulations. The grey areas show the respective quantiles 25%-75%, 10%-90% and 1%-99%. Left panel: Normal random effects with $\sigma^2 = 0.3^2$; Right panel: Mixture distribution for $V_t$ which has the same mean but a higher variance as before.