Noname manuscript No.
(will be inserted by the editor)

On COVID-19 Modelling

Extension of the article submitted to DMV

Robert Schaback

January 15, 2021 , **additions at the end, corrections in red, except for renumbering references and equations.**

Abstract This is an analysis of the COVID-19 pandemic by comparably simple mathematical and numerical methods. The final goal is to predict the peak of the epidemic outbreak per country with a reliable technique. The difference to other modelling approaches is to stay extremely close to the available data, using as few hypotheses and parameters as possible.

For the convenience of readers, the basic notions of modelling epidemics are collected first, focusing on the standard SIR model. Proofs of various properties of the model are included. But such models are not directly compatible with available data. Therefore a special variation of a SIR model is presented that directly works with the data provided by the Johns Hopkins University. It allows to monitor the registered part of the pandemic, but is unable to deal with the hidden part. To reconstruct data for the unregistered Infected, a second model uses current experimental values of the infection fatality rate and a data-driven estimation of a specific form of the recovery rate. All other ingredients are data-driven as well. This model allows predictions of infection peaks.

Various examples of predictions are provided for illustration. They show what countries have to face that are still expecting their infection peak. Running the model on earlier data shows how closely the predictions follow the transition from an uncontrolled outbreak to the mitigation situation by non-pharmaceutical interventions like contact restrictions.

Keywords Epidemiology, SIR model, ordinary differential equations

Mathematics Subject Classification (2010) 92D30 · 92D25 · 93C15 · 34A34

Prof. (em.) Dr. Robert Schaback
Institut für Numerische und Angewandte Mathematik,
Universität Göttingen, Lotzestraße 16-18, 37083 Göttingen
http://num.math.uni-goettingen.de/schaback
E-mail: schaback@math.uni-goettingen.de

1 Introduction and Overview

During an epidemic outbreak like COVID-19, everybody wants to know how hard the impact will be. In particular:

- What is the health risk for me, my family, our friends, the city, the country, and the world?
- Is the health system prepared properly?
- Should households fill up their reserves in time?

This is a situation that asks for mathematics, like in the old times when mathematicians were needed to predict floods or solstices. Such predictions should be based on data and arguments, and they should provide well-supported suggestions for what to do. To understand the process and to make predictions, it should be modelled, and the model should be computable. Then predictions will be possible, and reality will decide later whether the model and the predictions were useful. Many models are possible, and the approach presented here is just one of them. The specific goal is to stay as close as possible to the available data, but it turns out that the available data are not directly usable for the standard models that give the basic understanding. To this end, two extensions to the standard SIR model are developed that get closer to the available data and finally are able to make data-driven predictions.

The beginning is made in section 2 with an introduction to standard terms like *Basic Reproduction Number*, *Herd Immunity Threshold*, and *Doubling Time*, together with some critical remarks on their use in the media. These notions are based on the standard SIR model for epidemics that is treated in quite some detail, including proofs for most of the mathematical properties. Experts can skip over this completely. Readers interested in the predictions should jump right away to section 5. For simplicity, the presentation ignores all delay-related issues like *incubation period* and *serial interval*.

To bridge the gap between model and data, Section 3 describes the Johns Hopkins data source with its limitations and flaws, and then presents a variation of a SIR model that can be applied directly to the data. It allows to estimate basic parameters, including the Basic Reproduction Number. But since the Johns Hopkins data provide no information about the unregistered cases and the Susceptibles, the model cannot yield reliable predictions of peaks of epidemics.

Therefore section 4 combines the data-compatible model of section 3 with a SIR model dealing with the unknown Susceptibles and the unregistered Infectious. This needs extra parameters that must be extracted from the literature. The first is the *infection fatality rate*, as provided e.g. by an der Heiden/Buchholz [11], Streeck et al. [29], Verity et al. [31]. Section 4.3.1 pairs it with the *case fatality rate* and shows how the latter can be deduced from the Johns Hopkins data. Like in Bommer/Vollmer [2], their combination gives a detection rate for the confirmed cases.

Section 4.4 introduces the second additional parameter: a recovery rate that can be directly used in the model and estimated from the infection fatality rate and the observable case fatality and case death rates. However, this parameter is not needed for prediction, just for determination of the unknown variables from the known data as long as the latter are available.

Then section 5 combines all of this into a larger model that makes predictions under the assumption that there are no further changes to the parameters by political action. It estimates the parameters of a full SIR model from the available Johns-Hopkins data by the techniques of section 4, using two additional technical parameters: the number of days used backwards for estimation of constants, and the number of days in which recovery or death can be expected on average, for estimation of case fatality and recovery rates. This is where time delays enter, but not into the model, only into internal estimation procedures. After the data-driven estimation of these parameters, the prediction uses only the infection fatality rate. All other ingredients are derived from the Johns Hopkins data.

Results are presented in section 5. Given the large uncertainties in the Johns-Hopkins data, the predictions are rather plausible. However, reality will have the final word on this prediction model.

The paper closes with a summary and a list of open problems.

2 Classical SIR Modelling

This contains the basic notions for modelling epidemics, defined and explained in mathematical terms. In particular, there will be a rigid mathematical underpinning of what is precisely meant when media talk about

- *flattening the epidemic outbreak (mitigation)*,
- *basic reproduction number*,
- *Herd Immunity Threshold*, and
- *doubling time*,

pointing out certain abuses of these notions. This will not work without calculus, but things were kept as simple as possible. Readers from outside the mathematics community should take the opportunity to brush up their calculus knowledge. Experts should go over to section 3.

2.1 The Model

The simplest standard *SIR* model of epidemics, due to Kermack-McKendrick [16] in 1927 and easily retrievable from the Wikipedia [34], deals with three variables

Susceptible (S), Infectious (I), and Removed (R).

The Removed cannot infect anybody anymore, being either dead or immune. This is the viewpoint of bacteria or viruses. The difference between death and immunity of subjects is totally irrelevant for them: they cannot proliferate anymore in both cases. The SIR model cannot say anything about death rates of persons.

The Susceptible are not yet infected and not immune, while the Infectious can infect Susceptibles. Individuals move by infection from S to I, and by death or healing from I to R. The three classes S, I, and R are disjoint and add up to a fixed total population count $N = S + I + R$. All of these are ideally assumed to be smooth functions

of time t, and satisfy the differential equations

$$\begin{aligned} \dot{S} &= -\beta \frac{S}{N} I, \\ \dot{I} &= +\beta \frac{S}{N} I - \gamma I, \\ \dot{R} &= \gamma I. \end{aligned} \tag{1}$$

where the dot stands for the time derivative, and where β and γ are positive parameters. The product $\frac{S}{N}I$ models the probability that an Infectious meets a Susceptible and is actually infected.

Managing an SIR epidemic means *modifying* the constants β and γ. This is why one should see the parameters as control variables, and we shall treat them even as time series from section 3 on.

Note further that the Removed of the SIR model are not the Recovered of the Johns Hopkins data that we treat later, and the SIR model does not account for the Confirmed counted there. Similarly, there is no direct relation to the data published by the Robert Koch Institute. It is a major problem to match models with the available data, and we shall explain the latter to some detail in section 3. The inventors Kendrick and McKermack fitted their model already in 1927 [16] to data from the plague in Bombay 1905-1906.

2.2 Other Models

In many publications concerning COVID-19 (e.g. an der Heiden/Buchholz [11], Dandekar/Barbasthatis [3], De Brouwer et al. [4], Friston et al. [8], Khailaie et al. [17], Kucharski et al. [18], Maier/Brockmann [19]), the SIR model is extended by Exposed E that are infected, but not (yet) infectious. This introduces an additional parameter and would require dealing with a latency delay properly. We avoid this complication to keep the model as simple as possible. Note that there are extensions of SIR models with 14 to 21 parameters, e.g. Friston et al. [8], Giordano et al. [9], Khailaie et al. [17]. Fitting model parameters in the above papers is partially done numerically and partially by Bayesian approaches using Markov chain sampling of prior distributions. Here, we avoid fitting and time delays as far as possible.

Conceptually different are the agent-based model that is used by Ferguson et al. [7] for parameter estimation, and the approach of Mohring et al. [20] working consistently with time delays.

2.3 Simple Properties of the SIR Model

Since $\dot{N} = \dot{S} + \dot{I} + \dot{R} = 0$ holds in (1), the equation $N = S + I + R$ is kept valid at all times. The term $\beta \frac{S}{N} I$ moves Susceptibles to Infectious, while γI moves Infectious to Removed. Thus β represents an *infection rate* while the *removal rate* γ accounts for either healing or fatality after infection, i.e. immunity. Political decisions about

reducing contact probabilities will affect β, while γ resembles the balance between the medical aggressivity of the infection and the quality of the health care system.

As long as the Infectious I are positive, the Susceptibles S are decreasing, while the Removed R are increasing. Excluding the trivial case of zero Infectious from now on, the Removed and the Susceptible will be strictly monotonic. Therefore we can use them to re-parameterise the model at certain places.

The SIR model is not really dependent on the total population N. Moreover, if we scale time by $\tau := t \cdot \gamma$ and go over to *relative* quantities

$$\begin{aligned} s(\tau) &:= \frac{S(\tau/\gamma)}{N}, \\ r(\tau) &:= \frac{R(\tau/\gamma)}{N}, \\ i(\tau) &:= \frac{I(\tau/\gamma)}{N}, \end{aligned}$$

we get the new system

$$\begin{aligned} s'(\tau) = \frac{ds}{d\tau} &= -\frac{\beta}{\gamma} s(\tau) i(\tau) &&= -R_0 s(\tau) i(\tau) \\ i'(\tau) = \frac{di}{d\tau} &= \left(\frac{\beta}{\gamma} s(\tau) - 1\right) i(\tau) &&= (R_0 s(\tau) - 1)\, i(\tau) \\ r'(\tau) = \frac{dr}{d\tau} &= i(\tau) \end{aligned} \tag{2}$$

only containing the *Basic Reproduction Number*

$$R_0 := \frac{\beta}{\gamma} \tag{3}$$

that will turn out to be of central importance. Both β and γ vary under a change of time scale in (1), but the basic reproduction number is invariant. Physically, β and γ have the dimension $time^{-1}$, but $R_0 = \beta/\gamma$ and the new "time" parameter τ in (2) are dimensionless. Another interpretation of (2) is that after a time scale one can assume $\gamma = 1$ and $R_0 = \beta$. We call τ the *unit removal parameter*, because its unit can be seen as the average time needed to get removed, i.e. either dead or immune. We use a prime to denote derivatives with respect to τ. But in all later sections that make real-world interpretations, we have to use real time, and then we shall go back to (1).

A standard mathematical trick is to divide the first equation by the third to get

$$\begin{aligned} \frac{ds}{dr} &= -R_0 s, \\ s(r) &= s(r(0)) \exp(-R_0 (r - r(0))). \end{aligned} \tag{4}$$

We shall use (4) in section 2.11 to study the long-term behaviour of solutions. The introduction of (4) is a typical pitfall for mathematics: it is a nice theoretical simplification, but it obscures the most interesting practical aspect, in this case the fraction i of infectious persons in the population. The same holds for the simplification by setting $d\tau = \gamma \frac{I}{N} dt$ that is ignored here, leaving it to interested readers.

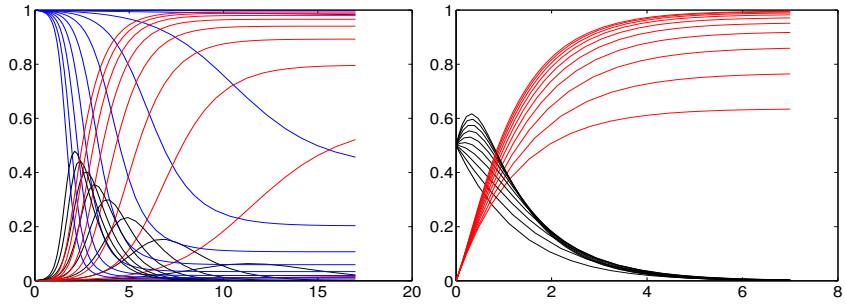

Fig. 1 Some typical SIR system solutions, relative to the total population. See the explanation in section 2.4. The peaked curves for the Infectious are "flattened" for small R_0.

2.4 Examples

Figure 1 shows a series of test runs of a SIR model. Recall that the relative Recovered r are increasing from zero, and the relative Susceptibles s are decreasing down from one. The relative Infectious i are in between and can possibly show a sharp peak that everybody tries to avoid. We shall deal with the mathematics of the peak in sections 2.8, 2.13, and 2.14, while the rest of the paper focuses on data-driven predictions of peaks. The Infectious are usually not covered by the media who tend to focus on the cumulative number of confirmed cases, containing the Removed.

In both plots we set $r(0) = 0, \gamma = 1$, and let $R_0 = \beta$ vary from 0.1 to 5. The difference between the figures lies in the initial value $i(0)$. Left, due to a realistically small $i(0) = 0.001$, one cannot see the decaying peak-less cases of i near startup for $R_0 < 1$, while the right-hand plot has $i(0) = 1/2$ and shows them. Decreasing $R_0 \searrow 1$ flattens the peaks of the Infectious i, and there is no peak for $R_0 \le 1$. Furthermore, one can observe that i always decays to zero, while s and r tend to fixed positive levels in the long run. The final level of r is particularly interesting because part of it is the total death toll. It decreases when R_0 decreases. We shall prove all of this later. When countries change parameters by administrative actions like a shutdown, they jump to a more flat i curve, e.g. at an intersection point.

From the system, one can also infer that r has an inflection point where i has its maximum, since $r'' = i'$. If only r would be observable, one could locate the peak of i via the inflection point of r. Finally, note that small initial values $i(0)$ of i delay the peak considerably, no matter how large R_0 is. We shall prove this in section 2.14.

2.5 Interpretation of the Basic Reproduction Number R_0

Media often say that R_0 gives the number of persons an average Infectious infects while being infectious. This is a rather mystical statement that needs underpinning. In the SIR system (1) the quantity

$$\frac{1}{\gamma} = \frac{I}{\dot{R}}$$

is a value that has the physical dimension of time. It describes the ratio between current Infectious and current newly Removed, and thus can be seen as the average time needed for an Infectious to get Removed, i.e. the average time that an Infectious can infect others. This is why we called the dimensionless $\tau = t \cdot \gamma$ the *unit removal parameter* in section 2.3. Correspondingly,

$$\dot{I} + \gamma I = \dot{I} + \dot{R} = \beta \frac{S}{N} I$$

are the newly Infected, and therefore

$$\frac{1}{\beta} \frac{N}{S} = \frac{I}{\dot{I} + \dot{R}}$$

can be seen as the time it needs for an average Infectious to generate a new Infectious. The ratio $R_t := \frac{\beta}{\gamma} \frac{S(t)}{N}$ then gives how many new Infectious can be generated by an Infectious while being infectious. This is the time-dependent *Reproduction Number*, but it is only close to R_0 if $S(t) \approx N$, i.e. at the start of an outbreak. A correct statement is that R_0 is the average number of infections an Infectious generates while being infectious, but within an unlimited supply of Susceptibles.

To let less new Infectious be generated, administrative actions try to change the parameters of the epidemic towards small R_0. We shall see that this is correct from a mathematical viewpoint as well, and we shall study the influence of R_0 to quite some detail.

The above interpretation of R_0 shows two major ways to make R_0 small: reducing the number of possibly infective contacts, and reducing the time an Infectious has to infect others. The second works by putting all infectious persons into strict quarantine, while first can be done by reducing contacts of all persons, even the Susceptibles, and reducing the infection probability for each contact, e.g. by wearing masks.

SIR-based models of the COVID-19 pandemics estimate R_0 between 2 and 6 during an uncontrolled outbreak (see e.g. the Robert Koch-Institute [23], De Brouwer et al. [4], Dehning et al. [6], and Maier/Brockmann [19]), while *non-pharmaceutical interventions* (NPI) bring R_0 below 1. We shall see examples in 3.3.2 and 5.2.

The use of the Basic Reproduction Number R_0 in the media suggests that large R_0 are generally serious, because each Infectious infects several people. This is only true at the beginning of an outbreak, because then there are enough Susceptibles. But it will turn out in section 2.8 that the Infectious will always finally go to zero, whatever the Basic Reproduction Number is. See Figure 1 as well.

2.6 Conditions for Outbreaks

The first interesting question in a beginning epidemic is:

Will there be a serious outbreak, or will the infection disappear quickly?

Therefore we first look at the initial conditions for the model. Since everything is invariant under an additive time *shift*, we can start at time 0, and since time *scales* are irrelevant to the problem at startup, we can use the simplified system (2).

The relative Infectious i in (2) do not increase right from the start if $\dot{I}(0) \leq 0$, i.e.

$$s(0) \leq \frac{1}{R_0}, \tag{5}$$

and then they decrease further since the Susceptibles s must decrease and

$$\frac{i(\tau)'}{i(\tau)} = (\log i(\tau))' = R_0 s(\tau) - 1 < R_0 s(0) - 1 \leq 0. \tag{6}$$

There is no outbreak, and this must occur for all initial conditions if $R_0 \leq 1$. But if $R_0 > 1$, the outbreak depends on the initial condition (5). Altogether, outbreaks are fully characterised by

$$1 > s(0) > \frac{1}{R_0}. \tag{7}$$

2.7 Herd Immunity Threshold

In connection with an outbreak, the *Herd Immunity Threshold*

$$HIT = 1 - \frac{1}{R_0}$$

is often mentioned. The background question is:

> If an uninfected population is threatened by an infection with Basic Reproduction Number R_0, what is the number of immune persons needed to prevent an outbreak right from the start?

In the idealised situation $i(0) = 0$ and $s(0) + r(0) = 1$,

$$r(0) = 1 - \frac{1}{R_0} = HIT$$

follows from (5) and (7) as the threshold between outbreak and decay for the relative Removed. This does not refer to a whole epidemic scenario. It is to be checked *before* anything happens, and useless within a developing epidemic, whatever the media say.

2.8 The Peak

In the outbreak case (7), the main questions are:

- When will the Infectious reach their maximum?
- How large will the maximal value be?

More generally, we ask for a time t_I or a unit removal parameter $\tau_I = \gamma t_I$ where the Infectious i are positive and do not change. Then we have

$$0 = \frac{di}{d\tau}(\tau_I) = (R_0 s(\tau_I) - 1) i(\tau_I), \tag{8}$$

and the monotonicity of s implies uniqueness of τ_I and

$$s(\tau_I) = \frac{1}{R_0}. \tag{9}$$

If i would increase without reaching a maximum in finite time, the first equation of (2) would imply that s goes exponentially to zero, but then there is a τ_I with (9), and (8) follows. Summarising, this proves that whenever there is an outbreak by (7), there is a unique maximum of the relative Infectious i that we call the *peak* from now on. Behind the peak, or apart from any outbreak situation, the Infectious must go exponentially to zero due to (6), because the Susceptibles continue to decrease, no matter how large R_0 is.

Determining the peak is theoretically difficult, and in practice it requires good estimates for β and γ. Mathematical results on the peak will be in sections 2.13 and 2.14, while data-driven predictions follow in section 5.2

In real life it is highly important to avoid the peak situation, and this can only be done by administrative measures that change β and γ in (1) to the situation $\beta < \gamma$. This is what management of epidemics is all about, provided that an epidemic follows the SIR model. We shall see how countries perform.

In the peak situation of (8) and (9), the fraction

$$1 - \frac{1}{R_0} = 1 - s(\tau_I) = r(\tau_I) + i(\tau_I) \geq i(\tau_I) \tag{10}$$

of the relative Non-Susceptible at the peak is exactly the Herd Immunity Threshold. Thus it is correct to say that if the Immune of a population are below the Herd Immunity Threshold at startup, and if the Basic Reproduction Number is larger than one, the sum of the Immune and the Infectious will rise up to the Herd Immunity Threshold and then the Infectious will decay. This is often stated imprecisely in the media.

2.9 Analysing the Outbreak

When an outbreak starts, almost everybody is susceptible, i.e. $s(0) \approx 1$, and then

$$i' = R_0 s - 1 \approx R_0 - 1$$

models an exponential outbreak with exponent $R_0 - 1 > 0$ in unit removal parametrisation, with a solution

$$i(\tau) \approx i(0) \exp((R_0 - 1)\tau).$$

If this is done in real time t and discrete time steps Δt, the system (1) yields

$$\frac{I(t + \Delta t)}{I(t)} \approx \exp((\beta - \gamma)\Delta t).$$

The severity of the outbreak in real time is not controlled by $R_0 = \beta/\gamma$, but rather by $\beta - \gamma$. Publishing single values $I(t)$ does not give any information about $\beta - \gamma$. Better is the ratio of two subsequent values

$$\frac{I(t_2)}{I(t_1)} \approx \exp((\beta - \gamma)(t_2 - t_1)), \tag{11}$$

and if this gets smaller over time, the outbreak gets less dramatic because $\beta - \gamma$ gets smaller. But (11) is by mo means a correct way to estimate R_0.

Therefore, really useful information about an outbreak must concern I, but should not consist of single values. Increments in percent are much better, because their logarithm is proportional to $\beta - \gamma$. However, it needs increments of increments to see whether administrative actions are successful by changing $\beta - \gamma$. This is what the media rarely provided during the outbreak. On the positive side, the severity of a future outbreak in unit removal parameterisation is described correctly by estimates of $R_0 > 1$, if these have a solid mathematical and experimental basis. All changes of R_0 should be carefully monitored.

2.10 Doubling Time

Another information used by media during an outbreak is the *doubling time*, i.e. how many days it takes until daily values double. It is $n\Delta t$ with the number n from

$$2 = \frac{I(t + n\Delta t)}{I(t)} \approx \exp((\beta - \gamma)n\Delta t) = (\exp((\beta - \gamma)\Delta t)^n$$

or

$$n = \frac{\log 2}{(\beta - \gamma)\Delta\tau},$$

i.e. it is inversely proportional to $\beta - \gamma$. If political action doubles the doubling time, if halves $\beta - \gamma$. If politicians do this repeatedly, they never reach $\beta < \gamma$, and they never escape an exponential outbreak if they do this any finite number of times. Extending the doubling time will never prevent a peak, it only postpones it and hopefully flattens it. When presenting a doubling time, media should always point out that this makes only sense during an exponential outbreak. And it is not related to the basic reproduction number $R_0 = \beta/\gamma$, but to the difference $\beta - \gamma$.

2.11 Long-term Behaviour

Aside from the peak, it is interesting to know the portions of the population that get either permanently removed (by death or immunity) or never come into contact with the infection. This concerns the long-term behaviour of the Removed and the Susceptibles. Figure 1 demonstrates how r and s level out under all circumstances shown, but is this always true, and what is the final ratio? And if one has additional information on the percentage of casualties within the Removed, what is the total death toll in the long run?

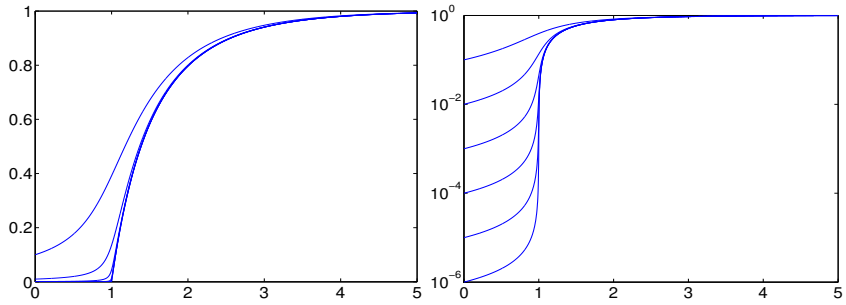

Fig. 2 The asymptotic level r_∞ of the relative Removed as a function of R_0 for $s(0) = 0.9, 0.99, 0.999$ etc. as curves from the top. Right: logarithmic scale.

Going back to (4), we get

$$s(r) = s(0)\exp(-R_0 r) \tag{12}$$

when assuming $r(0) = 0$ at startup. Since r is increasing, it has a limit $0 < r_\infty \leq 1$ for $\tau \to \infty$, and in this limit

$$s_\infty = s(0)\exp(-R_0 r_\infty) \tag{13}$$

holds, together with the condition $r_\infty + s_\infty = 1$, because there are no more Infectious. The transcendental equation

$$s(0)\exp(-R_0 r_\infty) = 1 - r_\infty \tag{14}$$

has a unique solution in $(0,1)$ dependent on $s(0) < 1$ and R_0. Therefore the Infectious always go to zero, but Susceptibles always remain. Then a new infection can always arise as soon as an infected person enters the sanitised population. The outbreak risk is dependent on the portion $s_\infty = 1 - r_\infty$ of the Susceptibles by (5). This illustrates the importance of vaccination, e.g. against measles or influenza.

To see how r_∞ and $s_\infty = 1 - r_\infty$ behave as functions of R_0 and $s(0)$, we solve the equation (14) by the Lambert W function to get

$$r_\infty = 1 + \frac{1}{R_0} W\left(-s(0)R_0\exp(-R_0)\right) \tag{15}$$

with a surprising behaviour. See Figure 2 for illustration. Left, the curves for unrealistically small initial values $s(0) = 0.9,\ 0.99$ and 0.999 for Susceptibles can still be distinguished from the more interesting curves below that coincide for all $s(0)$ closer to one and have a sharp turn at $R_0 = 1$. The logarithmic plot to the right shows that for $R_0 < 1$ the curves separate, and that it pays off significantly to have $R_0 < 1$ for $s(0)$ close to one.

This has some serious implications, if the model is correct for an epidemic situation. When politicians try to "flatten the curve" by bringing R_0 below 1 at some early time when the Susceptibles are still abundant, the asymptotic rate r_∞ of Removed will be *dramatically* smaller than for any other situation, because one stays left of the

sharp turn in Figure 2. This is particularly important if the rate of fatalities within the Removed is high.

Large values of R_0 lead to large relative values of Removed to Susceptible in the limit. The consequence is that systems with large R_0 have a dramatic outbreak and lead to a large portion of Removed. This is good news if the rate of fatalities within the Removed is low, but very bad news otherwise. When pressing R_0 below one, the risk of re-infection rises due to the larger portion of Susceptibles, but the deaths contained in the Removed are kept low.

The decay situation (5) implies that $s_\infty \le 1/R_0$ holds, and consequently

$$r_\infty = 1 - s_\infty \ge 1 - \frac{1}{R_0} = HIT.$$

Therefore the final rate of the Removed is not smaller than the Herd Immunity Threshold. This is good news for possible re-infections, but only if the death rate among the Removed is small enough.

2.12 Asymptotic Exponential Decay

If we go back to (6) for a unit removal parameter τ_D where i decreases, in an outbreak or not, we have $R_0 s_\infty \le R_0 s(\tau_D) < 1$ and then

$$i(\tau_D)\exp((R_0 s_\infty - 1)(\tau - \tau_D)) \le i(\tau) \le i(\tau_D)\exp((R_0 s(t_D) - 1)(\tau - \tau_D))$$

for all $\tau \ge \tau_D$. Therefore the exponential decay in unit removal parametrisation is not ruled by $R_0 - 1$ as in the outbreak case with $R_0 > 1$, but rather by $R_0 s_\infty - 1$. This also holds for large R_0 because s_∞ counteracts. The bell shapes of the peaked i curves are not symmetric with respect to the peak. Inserting (15), the relative Infectious always decay asymptotically exponentially like

$$\exp((R_0 s_\infty - 1)\tau) = \exp((W(-s(0)R_0\exp(-R_0)) - 1)\tau) \text{ for } \tau \to \infty \qquad (16)$$

with the Lambert W function. By MAPLE, the slowest decay arises for $R_0 = 1$.

2.13 Maximal Infectious at the Peak

At the peak of the Infectious i at τ_I in an outbreak (7) with $r(0) = 0$ we know

$$s(\tau_I) = \frac{1}{R_0} = s(r(\tau_I)) = s(0)\exp(-R_0 r(\tau_I))$$

from (9) and (4), and get the Removed at the peak as

$$r(\tau_I) = \frac{1}{R_0}\log(s(0)R_0). \qquad (17)$$

Then the exact value of the Infectious i at the peak is

$$i(\tau_I) = 1 - s(\tau_I) - r(\tau_I) = 1 - \frac{1}{R_0} - \frac{1}{R_0}\log(s(0)R_0), \qquad (18)$$

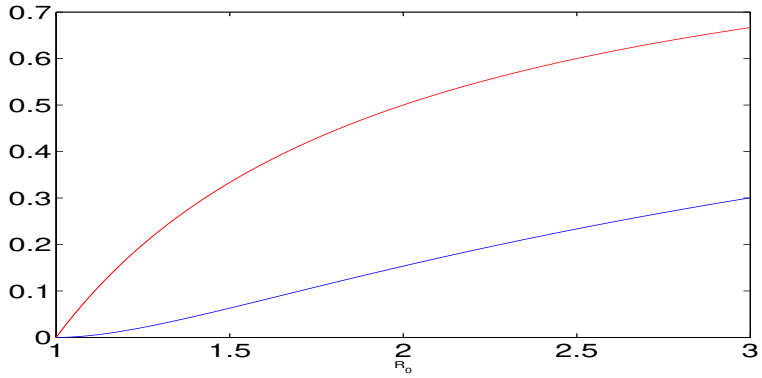

Fig. 3 The effect of R_0 on the peak value $i(\tau_i)$ of Infectious.

improving (10). Note that the log is positive due to the outbreak condition (7). It is remarkable that the *value* of i at the peak does not depend on initial conditions, while the next section proves that the *position* of the peak does.

For standard infections that have starting values $s(0) = S(0)/N$ very close to one, the maximal ratio of Infectious is

$$i(\tau_i) \approx 1 - \frac{1}{R_0} - \frac{1}{R_0}\log(R_0).$$

Figure 3 shows the behaviour of this function, as the lower curve. A value of $R_0 = 4$ leads to a maximum of more than 40% of the population infectious at a single time. If 5% need hospital care, a country needs hospital beds for 2% of the population around peak time. This disaster calls for mitigation by lowering R_0.

The upper curve leaves the log term out, i.e. it marks the rate (9) of the Susceptibles at the peak, and by (10) the difference is the rate $r(\tau_i)$ of the Recovered at the peak. It also marks the extreme case in (7) with $R_0 s(0) = 1$, i.e. having the smallest possible initial value of $s(0)$ for a given R_0 to generate an outbreak. Therefore all $s(0)$-dependent possibilities vary between the two curves.

2.14 Localising the Peak

Knowing now how large the peak is, we want to find out where it is. We write the unit removal parameter τ as a function of r by $\frac{d\tau}{dr} = (\frac{dr}{d\tau})^{-1} = \frac{1}{i}$ and integrate from $r = 0 = r(0)$ to $r = r(\tau_I)$ to get the peak position

$$\tau_I = \int_0^{r(\tau_I)} \frac{1}{i(r)} dr = \int_0^{\log(s(0)R_0)/R_0} \frac{1}{1 - r - s(0)\exp(-R_0 r)} dr \qquad (19)$$

as a nasty function of $s(0)$ and R_0, using (2), (12), and $1 = i(r) + s(r) + r$. To prove that the peak moves towards zero for both limits $R_0 \nearrow \infty$ and $R_0 \searrow 1$, we first observe

that $i \geq i(0)$ holds left of the peak. Then we use (17) to get

$$\tau_I \leq \frac{r(\tau_i)}{i(0)} = \frac{1}{i(0)R_0}\log(s(0)R_0) \leq \frac{1}{i(0)R_0}\log(R_0) \leq \frac{1}{e \cdot i(0)} \approx \frac{0.37}{i(0)} \tag{20}$$

by inserting the maximum of $\log(R_0)/R_0$ at e. The upper bound gets large when $i(0)$ gets small, a realistic case by Figures 1 and 8. This calls for a lower bound.

For fixed $s(0)$ and $i(0)$ there will be a maximal peak position for a rather specific R_0. A MAPLE-based analysis shows that $R_0(s(0)) = -W(-s(0)/e)^{-1}$ with Lambert's W function yields

$$\tau_I \geq \frac{0.3\,(1-i(0))}{\sqrt{i(0)}}. \tag{21}$$

Therefore the peak can indeed move arbitrarily far out for small $i(0)$ and large $s(0) = 1-i(0)$. There is not much leeway for smaller R_0 to bring the peak position to zero for large $s(0)$, namely $\frac{1}{s(0)} < R_0 < R_0(s(0))$. Both bounds for R_0 tend to one for $s(0) \to 1$.

The practical consequence is that keeping $R_0 > 1$ close to one by mitigation is no good idea, because the peak can move far into the future for realistically small $i(0)$, delaying the epidemic in an intolerable way. Countries should go for R_0 considerably smaller than one.

2.15 Turnaround Time

In a peak situation (7) one can consider the *turnaround parameter* τ_T at which the Infectious i come back to their starting value $i(0)$ behind the peak. At that point the population has accumulated more Removed, dead or immune. We calculate the integral

$$\int_0^\infty i(\tau)d\tau = \int_0^\infty r'(\tau)d\tau = r_\infty - r(0).$$

The rectangle of length τ_T and height $i(0)$ fits under the i curve, and therefore

$$i(0)\tau_T \leq r_\infty - r(0) \leq r_\infty \leq 1,$$

proving that the real turnaround *time* $t_T = \tau_T/\gamma$ has a fixed bound $t_T \leq r_\infty/(i(0)\gamma)$. From Figure 2 one can see that making R_0 smaller will decrease the bound via r_∞.

2.16 Estimating and Varying Parameters

If real-time data for the SIR model (1) were fully available, one could solve for

$$\begin{aligned} \gamma &= \frac{\dot R}{I}, \qquad\qquad b := \beta\frac{S}{N} = \frac{\dot I + \gamma I}{I} = \frac{\dot I + \dot R}{I}, \\ \beta &= \frac{N}{N-I-R}\cdot\frac{\dot I+\dot R}{I},\ R_0 = \frac{N}{N-I-R}\cdot\frac{\dot I+\dot R}{\dot R} = -\frac{N}{S}\frac{\dot S}{\dot R} = -\frac{1}{s}\frac{ds}{dr}, \end{aligned} \tag{22}$$

and we shall use this in section 3.3. The validity of a SIR model can be tested by checking whether the right-hand sides for β, γ and R_0 are roughly constant. If data

are sampled locally, e.g. before or after a peak, the above technique should determine the parameters for the global epidemic and be useful for either prediction or backward testing.

However, in pandemics like COVID-19, the parameters β and γ change over time by administrative action. This means that they should be considered as functions in the above equations, and then their changes may be used for conclusions about the influence of such actions. From this viewpoint, one can go back to the SIR model and consider β and γ as control functions that just describe the relation between the variables.

But the main argument against using (22) is that the data are rarely available. This is the concern of the next section.

3 Using Available Data

Now we confront the modelling of the previous section with available data. This is crucial for manœuvering countries through the epidemics (Sentker [26])[1]. From now on we have to work in real time and go back to (1) instead of all mathematical simplifications.

3.1 Johns Hopkins Data

We work with the COVID-19 data from the Johns Hopkins University at GitHub [10]. They are the only source that provides comparable data on a worldwide scale, namely

1. Confirmed (C) or *cumulative infected*
2. Dead (D), and
3. Recovered (R), i.e. alive and immune,

as cumulative integer valued time series for days from Jan. 22nd, 2020. All these values are absolute numbers, not relative to a total population. Note that the unconfirmed cases and the Susceptibles are not accessible at all, while the Confirmed contain the Dead and the Recovered of earlier days.

The media, in particular German TV, present COVID-19 data in a rather debatable way. When mentioning Johns Hopkins data, they provide C, D, and R separately without stating the most important figures, namely $I = C - D - R$, their change, and the change of their change. When mentioning data of the Infectious from the Robert Koch institute alongside, they do not say precisely that these are non-cumulative and should be compared to the $I = C - R - D$ data of the Johns Hopkins University. And, in most cases during the outbreak, they did not mention the change of the change. Quite like all other media.

We take the data as presented, but there are many well-known flaws. In particular, the values for specific days are partly belonging to previous days, due to delays in the chains of data transmission in different countries. This is why, at some points, we

[1] Original text in German, April 16th: *Schnelle Modelle, die dem Abgleich mit der Wirklichkeit standhalten, sind eine wichtige Voraussetzung, das Land politisch durch die Seuche zu steuern.*

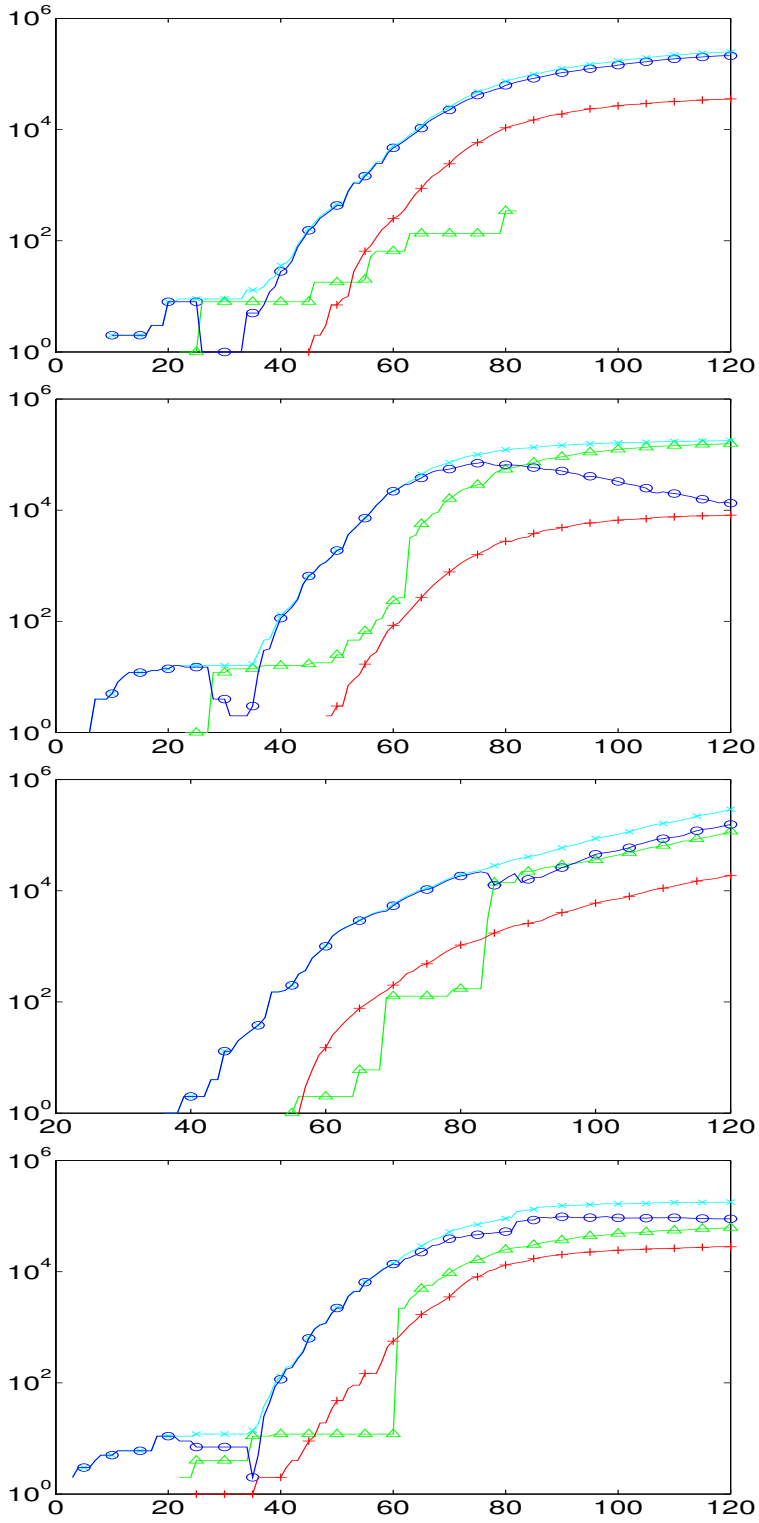

Fig. 4 Raw Johns Hopkins data in logarithmic presentation up to day 120, from top: UK, Germany, Brazil, and France. Markers X for Confirmed, O for Infectious, $\wedge$ for Recovered,$+$ for Deaths, not on all data points.

shall apply some conservative smoothing of the data. Finally, there are inconsistencies that possibly need data changes. In particular, there are countries like Germany who deliver data of Recovered in a very questionable way. The law in Germany did not enforce authorities to collect data of Recovered, and the United Kingdom did not report numbers of Dead and Recovered from places outside the National Health System, e.g. from Senior's retirement homes. Both strategies have changed somewhat in the meantime, as of early May, but the data still keep these flaws. See Figure 4 for examples.

We might assume that the Dead plus the Recovered of the Johns Hopkins data are the Removed of the SIR model, and that the Infectious $I = C - R - D$ of the Johns Hopkins data are the Infectious of the SIR model. But this is not strictly valid, because the Johns Hopkins data concern only registered cases.

On the other hand, one can take the radical viewpoint that facts are not interesting if they do not show up in the Johns Hopkins data. Except for the United Kingdom, the important figures concern COVID-19 casualties that are actually registered as such, others do not count, and serious cases needing hospitalisation or leading to death should not go unregistered. If they do in certain countries, using such data will not be of any help, unless other data sources are available.

An important point for what follows is that the data come as daily values. To make this compatible with differential equations, we shall replace derivatives by differences.

3.2 Examples

To get a first impression about the Johns Hopkins data, Figure 4 shows raw data up to day 120, May 21st. For better visibility, not all data points have markers. Here, and in all plots to follow, the x axis has the days after Jan. 22nd, 2020. It might be helpful to remember that day 100 is May 1st. The y axis is logarithmic, because then linearly increasing or decreasing parts in the figures correspond to exponentially increasing or decreasing numbers in the real data.

Many presentations in the media are non-logarithmic, and then all exponential outbreaks look similar. The most interesting data are the Infectious $I = C - R - D$ marked by O that show a peak or not, and the cumulative casualties D marked by $+$. The data for other countries tell similar stories and are suppressed.

One can see in Figure 4 that Germany has passed the peak of the Infectious, while France is roughly at the peak and the United States and Brazil are still in an exponential outbreak. The early figures, below day 40, are rather useless, but then an exponential outbreak is visible in all cases. This outbreak changes its slope due to political actions, and we shall analyse this later. See Dehning et al. [6] for a detailed early analysis of slope changes.

There are strange anomalies in the Recovered ($\wedge$ marker). France seems not to have delivered any data between days 40 and 58, Germany changed the data delivery policy between days 62 and 63, and the UK data for the Recovered are a mess. We shall avoid using data on the Recovered as much as possible.

It should be noted that the available medical results on the COVID-19 disease often state that Confirmed will die or survive after a more or less fixed number of days. This would imply that the curves marked $+$ for the Dead and the curves marked $\wedge$ for the Recovered should roughly follow the curves marked X for the Confirmed with a fixed but measurable delay. This is partially observable, but much less accurately for the Recovered.

3.3 The Johns Hopkins Data Model

We now define a model that works exclusively with the Johns Hopkins data, but comes close to a SIR model, without being able to use S. Since the SIR model does not distinguish between recoveries and deaths, we set in obvious notation

$$R_{SIR} \Leftrightarrow D_{JH} + R_{JH}$$

and let the Infectious be comparable, i.e.

$$I_{SIR} \Leftrightarrow I_{JH} := C_{JH} - D_{JH} - R_{JH}$$

which implies

$$(I+R)_{SIR} \Leftrightarrow C_{JH},$$

and we completely omit the Susceptibles. From now on, we shall drop the subscript JH when we use the Johns Hopkins data, but we shall use SIR when we go back to the SIR model.

Now we take (22) of section 2.16 and insert differences:

$$\begin{aligned}
\gamma &= \frac{\dot{R}_{SIR}}{I_{SIR}} \\
&\approx \frac{(D+R)_{n+1} - (D+R)_n}{I_n} =: \gamma_n \\
b := \beta \frac{S_{SIR}}{N} &= \frac{\dot{I}_{SIR} + \gamma I_{SIR}}{I_{SIR}} = \frac{\dot{I}_{SIR} + \dot{R}_{SIR}}{I_{SIR}}, \\
&\approx \frac{C_{n+1} - C_n}{I_n} =: b_n,
\end{aligned}$$

defining time series γ_n and b_n that model γ and $b = \beta \cdot S_{SIR}/N$ without knowing S_{SIR}. This is equivalent to the model

$$\begin{aligned}
C_{n+1} - C_n &= b_n I_n, \\
I_{n+1} - I_n &= b_n I_n - \gamma_n I_n = (b_n - \gamma_n) I_n, \\
(R+D)_{n+1} - (R+D)_n &= \gamma_n I_n
\end{aligned} \tag{23}$$

that maintains $C = I + R + D$, and we may call it a *Johns Hopkins Data Model*. It is very close to a SIR model if the time series b_n is not considered to be constant, but just an approximation of $\beta \cdot S_{SIR}/N$.

3.3.1 Estimating R

By brute force, one can take

$$r_n = \frac{b_n}{\gamma_n} = \frac{C_{n+1} - C_n}{R_{n+1} + D_{n+1} - R_n - D_n} \tag{24}$$

as a data-driven substitute for

$$\frac{\beta}{\gamma}\frac{S_{SIR}}{N} = R_0 \frac{S_{SIR}}{N}.$$

Then there is a rather simple observation:

If r_n is smaller than one, the Infectious decrease.

It follows using (24) via

$$\begin{aligned} I_{n+1} - I_n &= C_{n+1} - C_n - (R_{n+1} - R_n + D_{n+1} - D_n) \\ &= (r_n - 1)(R_{n+1} + D_{n+1} - R_n - D_n), \end{aligned}$$

but this is visible in the data anyway and not of much help.

Since r_n models $R_0 \frac{S_{SIR}}{N}$, it always underestimates R_0. This underestimation gets dramatic when it must be assumed that S_{SIR} gets seriously smaller than N.

At this point, it is not intended to forecast the epidemics. The focus is on extracting parameters from the Johns Hopkins data that relate to a background SIR-type model.

3.3.2 Example

Figure 5 shows $R_0 \frac{S_{SIR}}{N}$ estimates via r_n for the last four weeks before day 120, i.e. March 21st. Except for the United States and Brazil, all countries were more or less successful in pressing r_n below one. In all cases, S_{SIR}/N is too close to one to have any influence. The variation in r_n is not due to the decrease in S_{SIR}/N, but should rather be attributed to political action. As mentioned above, the estimates for R_0 by r_n are always optimistic.

For the figure, the raw Johns Hopkins data were smoothed by a double action of a $1/4, 1/2, 1/4$ filter on the logarithms of the data. This smoother keeps constants and linear sections of the logarithm invariant, i.e. it does not change local exponential behaviour. This smoothing was not applied to Figure 4. It was by far not strong enough to eliminate the apparent 7-day oscillations that are frequent in the Johns Hopkins data, see Figure 5, Data from the Robert Koch Institute in Germany have even stronger 7-day variations.

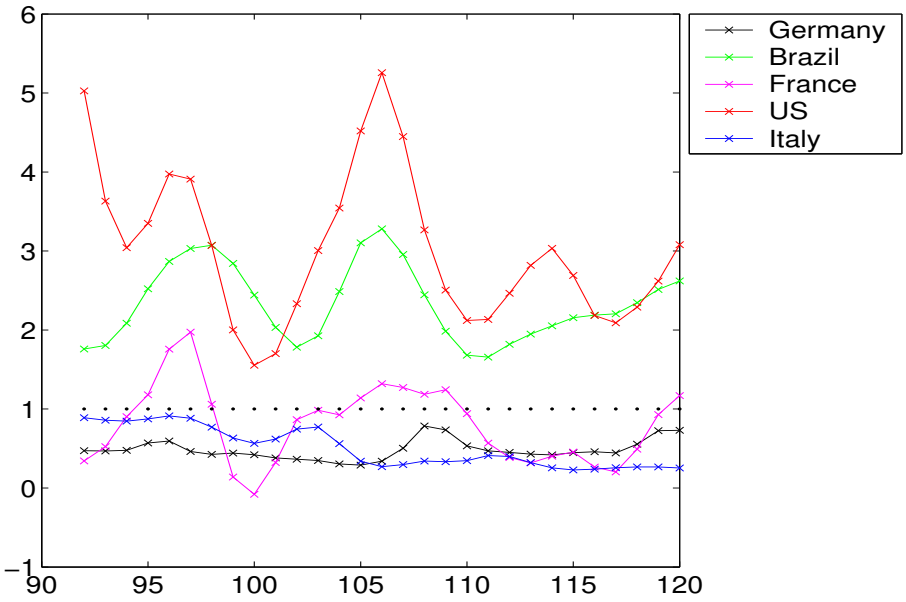

Fig. 5 Estimates of R_0 via the time series r_n up to day 120

3.3.3 Properties of the Model

As long as r_n is roughly constant, the above approach will always model an exponential outbreak or decay, but never a peak, because the difference equations are linear. It can only help the user to tell if there is a peak ahead or behind, depending on $r_n \approx R_0$ being larger or smaller than 1. If r_n is kept below one, the Confirmed Infectious will not increase, causing no new threats to the health system. Then the S/N factor will not decrease substantially, and a full SIR model is not necessary.

As long as countries keep r_n clearly below one, e.g. below 1/2, this would mean that $R_0 \approx r_n \frac{N}{S_{SIR}}$ stays below one if $S_{SIR} \geq N/2$, i.e. as long as the majority of the population has not been in contact with the SARS-CoV-2 virus. This is good news. But observing a small r_n can conceal a situation with a large R_0 if S_{SIR}/N is small. This is one reason why countries need to get a grip on the Susceptibles nationwide.

So far, the above argument cannot replace a SIR model. It only interprets the available data. However, monitoring the Johns Hopkins data in the above way will be very useful when it comes to evaluate the effectivity of certain measures taken by politicians. It will be highly interesting to see how the data of Figure 5 continue, in particular when countries relax their contact restrictions.

3.4 Extension Towards a SIR Model

For cases where one still has to expect $R_0 > 1$, e.g. US and Brazil on day 120 (see Figure 5), the challenge remains to predict a possible peak. Using the estimates from

the previous section is impossible, because they concern the sub-population of Confirmed and are systematically underestimating R_0. The "real" SIR model will have different parameters, a possibly large amount of undetected Infectious, and it needs the Susceptibles to model a peak and to make the r_n estimates realistic.

For an unrealistic scenario, consider *Total Registration*, i.e. all Infected are automatically confirmed. Then the Susceptibles in the Johns Hopkins model would be $S_n = N - C_n = N - I_n - R_n - D_n$. Now the estimate for R_0 must be corrected to

$$r_n \frac{N}{S_n} = r_n \frac{N}{N - C_n} = r_n \left(1 + \frac{C_n}{N - C_n}\right)$$

but this change will not be serious during an early outbreak.

If the time series $\beta_n = b_n \frac{N}{S_n} = b_n \frac{N}{N - C_n}$ for β and γ_n for γ are boldly used as predictors for β and γ in a SIR model, and if the model is started using $S_n = N - C_n = N - I_n - D_n - R_n$ in the discretised form

$$\begin{aligned} S_{n+1} - S_n &= -\beta \frac{S_n}{N} I_n, \\ I_{n+1} - I_n &= +\beta \frac{S_n}{N} I_n - \gamma I_n, \\ (R+D)_{n+1} - (R+D)_n &= -\gamma I_n, \end{aligned}$$

one gets a crude prediction of the peak in case $R_0 = \beta/\gamma > 1$.

Figure 6 shows results for two cases. The left plot shows the United States, using data from day 109 (May 10th) and estimating β and γ from the data one week before. The peak is predicted at day 473 (May 9th, 2021) with a total rate of 33% Infectious, i.e. about 124 million people. With an infection fatality rate of 0.5%, this means about 600,000 casualties in the two weeks around the peak. To see how crude the technique is, the second plot shows Germany using data up to day 75 (April 6th, 2020), i.e. before the peak, and the peak is predicted at day 230 (Sept. 8th, 2020) with about 16% Infected. This would imply about 65,000 casualties around the peak. At day 75, R_0 was estimated at 2.01, but a few days later the estimate went below 1 (Figure 5) by political intervention changing b_n considerably. See Figure 10 for a much better prediction using data only up to day 67.

4 Extended SIR Model

To get closer to reality, one should combine the data-oriented Johns Hopkins Data Model with a SIR model that accounts for what happens outside of the Confirmed. We introduce the time series

- S for the Susceptibles like in the SIR model,
- M for the Infectious, not yet confirmed, (M standing for *mysterious*),
- H for the unconfirmed Recovered (H standing for *healed*).

This implies that all deaths occur within the Confirmed, though this is a highly debatable issue. It assumes that persons with serious symptoms get confirmed, and nobody dies of COVID-19 without prior confirmation.

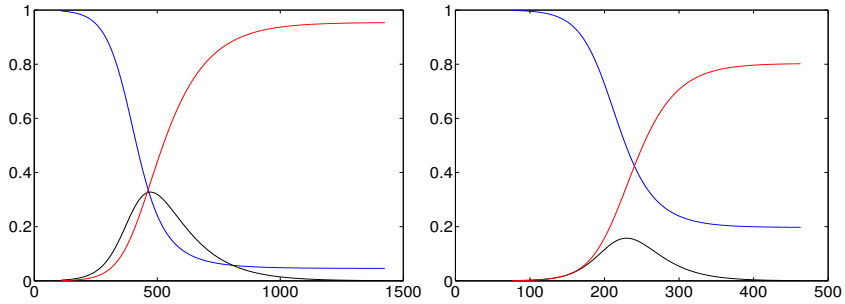

Fig. 6 Brute force SIR modelling for US and Germany using last week's data, at days 109 and 75, with $R_0 = 3.22$ and $R_0 = 2.01$, respectively.

4.1 The Hidden Model

The Removed from the viewpoint of a global SIR model including H and M are $H+C$, and thus the SIR model is

$$\begin{aligned} S_{n+1} - S_n &= -\beta \frac{S_n}{N} M_n, \\ M_{n+1} - M_n &= \beta \frac{S_n}{N} M_n - \gamma M_n, \\ (H+C)_{n+1} - (H+C)_n &= \gamma M_n. \end{aligned} \tag{25}$$

To run this *hidden* model with constant $N = S+M+H+C$, one needs initial values and good estimates for β and γ, which are not the ones of the Johns Hopkins Data Model of section 3.3. We need other ways to get them.

4.2 The Observable Model

The Johns Hopkins variables D and R are linked to the hidden model via $C = I - R - D$. They follow an *observable* model

$$\begin{aligned} I_n &= C_n - R_n - D_n, \\ D_{n+1} - D_n &= \gamma_{iCD} I_n, \\ R_{n+1} - R_n &= \gamma_{iCR} I_n \end{aligned} \tag{26}$$

with *instantaneous case death and recovery rates* γ_{iCD} and γ_{iCR} for the Confirmed Infectious. These rates can be estimated separately from the available Johns Hopkins data, and we shall do this below. We call these rates *instantaneous*, because they artificially attribute the new deaths or recoveries at day $n+1$ to the Infectious of the previous day, not of earlier days. They are *case* rates, because they concern the Confirmed. The difference between standard and instantaneous case rates will be treated in sections 4.3.1 and 4.3.2.

The observable model is coupled to the hidden model only by C_n. Any data-driven C_n from the observable model can be used to enter the $H+C$ variable of the

hidden model, but in an unknown ratio. Conversely, any version of the hidden model produces $H+C$ values that do not determine the C part. Summarising, there is no way to fit the hidden model to the data without additional assumptions.

Various possibilities were tried to connect the Hidden to the Observable. Two will be presented now.

4.3 Fatality Rates

4.3.1 Infection Fatality Rate

Recall that the parameter γ_{iCD} in the observable model (26) relates case fatalities to the confirmed Infectious of the previous day. In contrast to this, the *infection fatality rate* in the standard literature, denoted by γ_{IF} here, is relating to the infection directly, independent of the confirmation, and gives the probability to die of COVID-19 after infection with the SARS-CoV-2 virus, whatever the delay between infection and death is. It was estimated as $\gamma_{IF} = 0.56\%$ by an der Heiden/Buchholz [11] and 0.66% by Verity et al. [31], but specialised for China. Recent data of Streeck et. al. [29] gives a value of 0.36% for the Heinsberg population in Germany. For the UK, Ferguson et al. [7] arrive at 0.9%. We shall later use 0.5% for our predictions. But it is very desirable to get more information on infection fatality rates, in particular for different countries. So far, we use a single value globally.

The idea to use the infection fatality rate for information about the hidden system comes from Bommer/Vollmer [2]. The infection fatality rate will be used below in (31) and (33) together with case fatality rates that we consider next.

4.3.2 Estimation of Case Fatality Rates

We now focus on probabilities to die either after an infection or after confirmation of an infection. The first is the infection fatality rate given in the literature, but what is the latter, the *case fatality rate* γ_{CF} when using the Johns Hopkins data? It is clearly not the γ_{iCD} in (26), giving the ratio of new deaths at day $n+1$ as a fraction of the confirmed Infectious at day n. The deaths at day $n+1$ must be assigned to various earlier days instead.

Case fatality rates in the literature vary strongly, and they are country-dependent. Countries have different ways to detect cases, and because the mortality is age-dependent, different age structures will have a serious influence. The Robert-Koch-Institute [23] mentions 10.5% for Europe and 4.6% for Germany, while De Brouwer et al. [4] has 10.0% for Italy, 4.0% for China, 6.0% for Spain, and 4.3% worldwide. According to Streeck et al.[29], the current estimate of the case fatality rate in Germany by the World Health Organization (WHO) is between 2.2% and 3.4%.

We cannot clean up these inconsistencies. Instead, we now describe a way to estimate case fatality rates per country from the Johns Hopkins data. The basic idealistic assumption is that COVID-19 diseases end after k days from confirmation with either death or recovery. Let us call this the *k-day rule*. Suggested values for k start from 14 days for mild cases (an der Heiden/Buchholz [11] WHO [33]) and go up to 30

days, composed of an incubation time of about 5 days and various values between 11 and 25 days for hospitalisation, depending on the amount of intensive care (an der Heiden/Buchholz [11], Robert Koch-Institut [23], Verity et al. [31], Mohring et al. [20]).

Following Schaback [25], one can estimate the probability to survive on day $k+1$ after confirmation, and this works in a stable way per country, based only on C and D, not on the unstable R data. In [25] this approach was used to produce R values that comply with the k-day rule, but here we use it for estimating the case fatality.

The basic argument lets the new Confirmed $C_n - C_{n-1}$ at day n enter into the new deaths $D_{n+1} - D_n$ at day $n+1$ with probability $p_1 =: q_1$, into $D_{n+2} - D_{n+1}$ with probability $p_2(1-p_1) =: q_2$ and so on. The rest enters into the new Recovered at day $n+k$ with probability q_{k+1} if we set $p_{k+1} = 1$ and define

$$q_i = p_i \prod_{j=1}^{i-1}(1-p_j),\ 1 \le i \le k+1. \tag{27}$$

Then the estimated case fatality rate is $1-q_{k+1}$, while the case recovery rate is q_{k+1}. Therefore the technique of [25] performs a fit

$$D_n - D_{n-1} \approx \sum_{i=1}^{k} q_i (C_{n-i} - C_{n-i-1}), \tag{28}$$

over all nonnegative $q_1, \ldots, q_k$ with sum bounded by one and connected to the p_i by (27). It assigns all new deaths at day n to previous new infections on previous days in a hopefully consistent way, minimising the error in the above formula under variation of the probabilities p_i to die on day i after confirmation, and it delivers case fatality and case recovery rates per country. It formally assigns all recoveries to day $k+1$ after confirmation. Before that day, a living Confirmed cannot be declared to be recovered.

At this point, there is a hidden assumption. The change $C_{n+1} - C_n$ to the Confirmed is understood as the number of new registered infections, i.e. it is treated like $I_{n+1} - I_n$, disregarding short-time death or recovery. But replacing $C_{n-i} - C_{n-i-1}$ by $I_{n-i} - I_{n-i-1}$ in (28) would connect a cumulative function to a non-cumulative function. Furthermore, this requires the unsafe data of the Recovered.

In fact, the estimation via the fit (28) is unexpectedly reliable, provided one looks at $1-q_{k+1}$ or q_{k+1}, not at single probabilities p_j, and if sufficiently many n are used. This follows from a series of experiments that we do not document fully here, except for Figure 7. In [25], data for $2k$ days backwards were used for the estimation, and results did not change much when more or less data were used or when k was modified. Here, the range $7 \le k \le 21$ was tested, and backlogs of up to 50 days from day 109. See Figure 7 below for an example. It is typical here and for many other cases that a value of $k = 14$ performs well, with a backlog of $2k = 28$ days for the fit in (28). Using larger k needs a larger backlog, but then the estimation is not time-local enough to produce up-to-date estimates, because outdated values are used. Figure 7 shows the variation of the case fatality rate estimation when k and the backlog are varied. The rates usually do not vary much and have plateaus for $k \ge 14$,

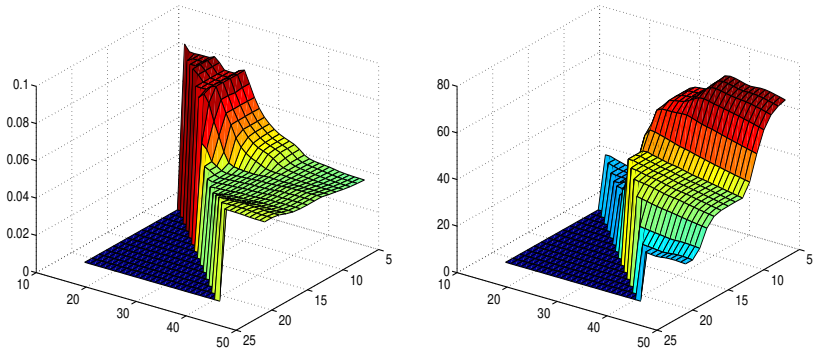

Fig. 7 Left: case fatality rate for Germany based on data at day 109, as functions of k (right axis) and the data backlog $B \geq 2k$ (left axis). Right: Root mean-square error for (28).

Country	Death rate	Detection rate
Germany	0.047	0.106
Brazil	0.094	0.053
Italy	0.138	0.036
Spain	0.085	0.059
Sweden	0.157	0.032
Austria	0.052	0.096
France	0.122	0.041
UK	0.145	0.035
US	0.067	0.075

Table 1 Case fatality and detection rates, estimated on day 109 using the 14-day rule and a backlog of 28 days.

but of course the errors decrease when k is taken larger, because there are more days to assign deaths to.

See the first column of Table 1 for estimates of case fatality rates for different countries, calculated on day 109 (May 10th) for $k = 14$ and a backlog of 28 days. They comply with the values from the literature cited above. Their interpretation depends strongly on the strategy for confirmation. In particular, they are high when only serious cases are confirmed, e.g. cases that need hospital care. If many more people are tested, confirmations will contain plenty of much less serious cases, and then the case fatality rates are low.

The instantaneous case death rate γ_{iCD} of (26) for the Johns Hopkins data comes out around 0.004 for Germany on day 109 by direct inspection of the data via

$$\gamma_{iCD} \approx \frac{D_{n+1} - D_n}{I_n}, \tag{29}$$

while the Case Fatality Rate γ_{CF} in Table 1 is about 0.047. The deaths have to be attributed to different days using the k-day rule, they cannot easily be assigned to the previous day without making the rate smaller.

4.3.3 The Detection Rate

A simple way to understand the quotient $\frac{\gamma_{IF}}{\gamma_{CF}}$ of the infection fatality rate γ_{IF} and the case fatality rate γ_{CF} as a *detection rate* is to ask for the probability $p(C)$ for Confirmation. If the probability to die after Confirmation is γ_{CF}, and if there are no deaths outside confirmation, then

$$p(D) = p(C) \cdot p(D|C),$$

by conditional probabilities, and

$$p(C) = \frac{p(D)}{p(D|C)} = \frac{\gamma_{IF}}{\gamma_{CF}}. \tag{30}$$

See the second column of Table 1, prepared for $\gamma_{IF} = 0.005$. The rate depends on good estimates of the infection fatality rate, and the new value 0.0036 by Streeck et al. [29] will decrease the detection rate for Germany from 10.6% to 7.7% for the Heinsberg subpopulation.

All of this is comparable to the findings of Bommer/Vollmer [2] and uses the basic idea from there, but with a somewhat different technique and different results. There, the values were 7% for March 23rd and 9% for March 30th, while Mohring et al. [20] assume 20% on April 29th.

4.3.4 Using Fatality Rates for the Hidden Model

If the case fatality rates γ_{CF} of Table 1 are used with a known infection fatality rate γ_{IF}, one should obtain an estimate of the total Infectious. If the formula (28) is written as

$$\sum_{i=1}^{k} q_i(C_{n-i} - C_{n-i-1}) \approx D_n - D_{n-1} \approx \sum_{i=1}^{k} \tilde{q}_i(S_{n-i-1} - S_{n-i})$$

in terms of the previous new infections $S_{n-i-1} - S_{n-i}$ in terms of Susceptibles with daily infection fatality probabilities $\tilde{q}_i$, one should maintain

$$\gamma_{CF} = \sum_{i=1}^{k} q_i \text{ and } \gamma_{IF} = \sum_{i=1}^{k} \tilde{q}_i,$$

and this works by setting

$$C_n - C_{n-1} = \frac{\gamma_{IF}}{\gamma_{CF}}(S_{n-1} - S_n) \tag{31}$$

in general, without using the unstable p_i. This is the first connection of the Observable to the Hidden, namely C to S. Like in the discussion following (28) one can argue to use M instead of S in (31), but this would again connect a cumulative variable to a non-cumulative one.

4.3.5 Local Estimation of Fatality Rates

Because politicians change testing strategies and the parameters β and γ, the estimation of the Case Fatality Rate should be made locally, not globally. Using the experience of Schaback [25] and section 4.3.2, we shall use a fixed $k = 14$ for the k-day rule and data for a fixed backlog of $2k$ days. Then the formula (31) has γ_{CF} varying with n as far as Johns Hopkins data are available.

4.4 Recovery Rates

We need another parameter to connect the hidden to the observable model. There are many choices, and after some failures we selected the constant γ_{iIR} in a model equation

$$H_{n+1} - H_n = \gamma_{iIR} M_n.$$

Following what was mentioned about *instantaneous* rates in section 4.2, γ_{iIR} is an *instantaneous Infection Recovery Rate*, relating the new unregistered Recovered to the unregistered Infections the day before.

4.4.1 Estimation of Recovery Rates

A good value of γ_{iIR} can come out of a field experiment that produces time series for M and H, i.e. for unregistered Infectious and unregistered Recovered. Then the instantaneous Infection Recovery rate γ_{iIR} can be obtained directly by

$$\frac{H_{n+1} - H_n}{M_n} \approx \gamma_{iIR}.$$

The Infection Recovery rate $\gamma_{IR} = 1 - \gamma_{IF}$ does not help, because we need an instantaneous rate that has no interpretation as a probability.

With the risk of using unstable data of the Recovered, we can look at the instantaneous Case Recovery rate

$$\frac{R_{n+1} - R_n}{I_n} \approx \gamma_{iCR} \tag{32}$$

that is available from the Johns Hopkins data, and comes out experimentally to be rather stable, provided that countries have useful data for the Recovered. Otherwise, we have to use the technique of Schaback [25] for estimating them. The rate γ_{iIR} must be larger than γ_{iCR} because we now are not in the subpopulation of the Confirmed, and nobody can die without going first into the population of the Confirmed. As long as no better data are available, we shall use the formula

$$\gamma_{iIR} = \frac{1 - \gamma_{IF}}{1 - \gamma_{CF}} \gamma_{iCR} = \frac{\gamma_{IR}}{\gamma_{CR}} \gamma_{iCR} = \frac{\gamma_{iCR}}{\gamma_{CR}} \gamma_{IR} \tag{33}$$

that implements two meaningful arguments:

1. the value γ_{iCR} is increased by the ratio $\frac{\gamma_{IR}}{\gamma_{CR}}$ of Recovered probabilities for the Infected and the Confirmed,

2. the value γ_{IR} is multiplied by a factor $\frac{\gamma_{iCR}}{\gamma_{CR}}$ for transition to immediate rates, and this factor is the transition factor for the Confirmed Recovered.

The above strategy is debatable and may be the weakest point of this approach. However, others turned out to be worse, mainly due to instability of results. On the positive side, the final prediction will not need it, see (38) below. It enters only the intermediate step when S, M, and H are calculated in the time range of the available Johns Hopkins data, see (34) in section 4.5. And, finally, there is hope that there will be field experiments that yield reliable values directly.

4.4.2 Practical Approximation of Recovery Rates

In (33) the rate γ_{IR} is fixed, and the rate γ_{CR} is determined locally via section 4.3.5. The rate γ_{iCR} follows from the time series

$$\frac{R_{n+1}-R_n}{I_n} \approx \gamma_{iCR}$$

as in (26). This works for countries that provide useful data for the Recovered. In that case, and in others to follow below, we can take the time series itself as long as we have data. For prediction, we estimate the constant from the time series using a fixed backlog of m days from the current day, i.e. we take the mean of the last $m+1$ values. Since many data have a weekly oscillation, due to data being not properly delivered during weekends, the backlog should not be less than 7.

But for certain countries, like the United Kingdom, the data for the Recovered are useless. In such cases, we employ the technique of Schaback [25] to estimate the Recovered using the k-day rule and a backlog of $2k$ days, like in section 4.3.5 for the case fatality rates.

4.5 Model Calibration

We now have everything to run the hidden model, but we do it first for days with Johns Hopkins data, delaying predictions to section 5. This is a *calibration step* that leads to estimations of S, M, and H from the observed data of the Johns Hopkins source, without any need for sophisticated fitting algorithms. With the parameters from above, we use the new relations

$$\begin{aligned} C_{n+1}-C_n &= \frac{\gamma_{IF}}{\gamma_{CF}}(S_n-S_{n+1}), \\ H_{n+1}-H_n &= \gamma_{iIR}M_n \end{aligned} \tag{34}$$

in a specific way. We set up the second model equation in (25) for M as

$$\begin{aligned} M_{n+1}-M_n &= S_n-S_{n+1}-\gamma_n M_n \\ &= \frac{\gamma_{CF}}{\gamma_{IF}}(C_{n+1}-C_n)-\gamma_n M_n \\ &= \frac{\gamma_{CF}}{\gamma_{IF}}(C_{n+1}-C_n)-(C_{n+1}-C_n+H_{n+1}-H_n) \\ &= \left(\frac{\gamma_{CF}}{\gamma_{IF}}-1\right)(C_{n+1}-C_n)-\gamma_{iIR}M_n \end{aligned} \tag{35}$$

that can be solved if an initial value $\tilde{M}_0$ is prescribed. Then (34) is run to produce the S_n and H_n, with starting values that we describe in section 4.5.1. If β_n and γ_n are calculated by

$$\begin{aligned} \beta_n \frac{S_n}{N} M_n &= S_n - S_{n+1}, \\ \gamma_n M_n &= C_{n+1} - C_n + H_{n+1} - H_n, \end{aligned} \tag{36}$$

respectively, the balance equation $N = S + M + H + C$ follows from (35) and (36).

4.5.1 Starting Values

Since the populations are large, the starting values for S are not important. Beginning at the full population N from a very early day, the S values are calculated from (34) first, just to get values S_j for actually starting at later days.

Then the first day j is taken where C_j is at least 10, and k days later the start value for H is set as

$$H_{j+k} = C_j \frac{\gamma_{CF}}{\gamma_{IF}} \tag{37}$$

using the k-day rule with $k = 14$. This divides the $C_j > I_j$ value by the detection rate, i.e. roughly all estimated undetected Infectious at time j are assumed to be healed k days later, i.e. at day $j + k$. Then the starting value for M_{j+k} is calculated via the balance equation $N = S + M + H + C$ from the S_{j+k} value calculated by the previous paragraph. Finally, the calibration starts at day $j + k$ by the above formulae. Unfortunately, this is a serious limit preventing application to very short time series.

The starting value for H is irrelevant for H itself, because only differences enter into the model, but it determines the starting value for M due to the balance equation. Anyway, it turns out experimentally that the starting values do not matter much, if the model is started early. The hidden model (25) depends much more strongly on C than on the starting values.

Figure 10 contains a wide variation of the starting value (37) for H at the starting point, by multipliers between 1/32 and 32. This has hardly any effect on the results, the lines getting somewhat thicker. The variation in starting values get more visible in other cases, see the right-hand plot in Figure 10 for the United States. But the influence on predictions is negligible.

4.5.2 Examples

The figures to follow in section 5.2 show the original Johns Hopkins data together with the hidden variables S, M, and H that are calculated by the above technique. The calibration runs up to the vertical line where predictions start. Note that the only ingredients beside the Johns Hopkins data are the number $k = 14$ for the k-day rule, the Infection Fatality Rate γ_{IF} from the literature, equations (34), and the backlog of $m = 7$ days for estimation of constants from time series.

5 Predictions using the Full Model

To let the combined model predict the future, or to check what it would have predicted if used at an earlier day, we take the calibrated model of the previous sections up to a day n and use the values $S_n, M_n, H_n, C_n, I_n, R_n$ and D_n for starting the prediction. With the variable $HC := H + C$, we use the recursion

$$\begin{aligned} S_{i+1} &= S_i - \beta \frac{S_i}{N} M_i, \\ M_{i+1} &= M_i + \beta \frac{S_i}{N} M_i - \gamma M_i, \\ HC_{i+1} &= HC_i + \gamma M_i, \\ C_{i+1} &= C_i + \gamma_{IF}(S_i - S_{i+1})/\gamma_{CF}, \\ R_{i+1} &= R_i + \gamma_{iCR} I_i, \\ D_{i+1} &= D_i + \gamma_{iCD} I_i, \\ I_{i+1} &= C_{i+1} - R_{i+1} - D_{i+1}, \\ H_{i+1} &= HC_{i+1} - C_{i+1}. \end{aligned} \tag{38}$$

This needs fixed values of β and γ that we estimate from the time series for β_n and γ_n by using a backlog of 7 days, following Section 4.5. The instantaneous rates γ_{iCR} and γ_{iCD} can be calculated via their time series, as in (32) and (29), using the same backlog. We do this at the starting point of the prediction, and then the model runs in a *no political change* mode. Examples will follow in section 5.2.

5.1 Properties of the Full Model

The first part of the full model (38) is a standard SIR model for the variables S, M and $H + C$, and inherits the properties of these as described in section 2. It does not use the γ_{IIR} parameter of the second equation in (34), and it uses the first the other way round, now determining C from S, not S from C.

The balances $N = S + M + H + C$ and $C = I + D + R$ are maintained automatically, and the time series for $S, C, R, H + C$, and D stay monotonic as long as M and I are non-negative. To check the monotonicity of H, consider

$$\begin{aligned} H_{i+1} - H_i &= HC_{i+1} - HC_i - C_{i+1} + C_i \\ &= \gamma M_i - \frac{\gamma_{IF}}{\gamma_{CF}}(S_i - S_{i+1}) \\ &= \left(\gamma - \beta \frac{\gamma_{IF}}{\gamma_{CF}} \frac{S_i}{N}\right) M_i. \end{aligned}$$

The bracket is positive if

$$R_0 = \frac{\beta}{\gamma} < \frac{\gamma_{CF}}{\gamma_{IF}} \frac{N}{S_i} \geq \frac{\gamma_{CF}}{\gamma_{IF}},$$

which is enough for practical purposes as long as detection rates $\frac{\gamma_{IF}}{\gamma_{CF}}$ are low and R_0 is not excessively large. Anyway, H should be monitored.

The slopes of S and C are always connected by (31), and those of R and D are connected by

$$R_{i+1} - R_i = \frac{\gamma_{iCR}}{\gamma_{iCD}}(D_{i+1} - D_i) \tag{39}$$

in the prediction part. But the figures below will show logarithms, and therefore the slope parallelism will not be visible.

By section 2.11, the hidden Infectious M will always go to zero, and the variables S and $H+C$ will level out in the long run. Since C is increasing, it must level out as well, and I must level out because R and D do. But due to the equations for R and D, the final level of I must be zero.

The asymptotic levels of S and $H+C$ follow from 2.11, but not the interesting level of D, the total death toll. If the prediction is started at day n, then

$$R_\infty - R_n = \frac{\gamma_{iCR}}{\gamma_{iCD}}(D_\infty - D_n),$$

obtained by summation of (39), connects the asymptotic deaths and confirmed recoveries. From the connection of S and C we likewise get

$$C_\infty - C_n = \frac{\gamma_{iIF}}{\gamma_{CF}}(S_n - S_\infty).$$

With $C_\infty = R_\infty + D_\infty$ we now have three independent equations for the unknowns C_∞, D_∞, R_∞. Because the theory of Section 2.11 yields S_∞ and $H_\infty + C_\infty$ in terms of β and γ, we know S_∞ and can get H_∞ from C_∞. But if the simulation is run long enough, one can easily read the asymptotic values off the plots.

5.2 Examples of Predictions

Figure 8 shows predictions on day 122, May 23rd, for Germany, Brazil, France, and USA, from the top. The plots for countries behind their peak are rather similar to those for Germany and France. The other two countries are selected because they still have to face their peak, if no action is taken to change the parameters.

The plots show that Germany can expect to get away with no more than 10000 casualties in the long run, while Brazil goes for a peak of about 20 million hidden Infectious in fall 2020 (M, symbol $\Box$) and a final death toll of about 1 million (D, symbol $+$). The United States would have to face a peak of hidden Infectious of about 25 million in mid-January 2021, and more than 1 million COVID-19 deaths in October 2021, and still rising. But of course, these predictions assume that reality follows the model and that there are no parameter changes by political action.

The estimated R_0 values are 0.65, 2.19, 0.42, and 1.75, respectively. Note that these are not directly comparable to Figure 5, because they are the fitted constants to the backlog of a week, and using (36) instead of (24), avoiding the systematic underestimation of the latter. The hidden M and H (symbols $\Box$ and $\diamond$) follow roughly the observable I and C (symbols O and x), but with a factor due to the detection rate that is different between countries, see Table 1. To enhance visibility, not all data points in the plots are marked with symbols. The C, R, I and D data left of the vertical

line are the original Johns Hopkins data. The S, M, H data there are calculated by section 4, while to the right the data are predictions for all variables by the full model (38).

All test runs were made for the infection fatality rate $\gamma_{IF} = 0.005$, the delay $k = 14$ for estimating case fatalities, and a backlog of 7 days when estimating constants out of recent values of time series. The choice $\gamma_{IF} = 0.005$ is somewhat between 0.56% from an der Heiden/Buchholz [11], 0.66% from Verity et al. [31], and 0.36% from Streeck et al. [29]. New information on infection fatality rates should be included as soon as they are available, and if possible per country, not global.

When used within estimation routines, the Johns Hopkins data were smoothed by a double application of the 1/4, 1/2, 1/4 filter on the logarithms, like for Figure 5. But the plots show the original Johns Hopkins and prediction data.

5.3 Evaluation of Predictions

To evaluate the prediction quality, one should go back and start the predictions on earlier days, to compare with what happened later. Figure 9 shows over-plots of predictions for days 94, 108, and 122, each a fortnight apart, though there may be parameter changes in the meantime. The starting points of the predictions are marked by vertical lines again. For better visibility, only the death count D (symbol $+$) and the two non-cumulative variables M and I for the hidden and confirmed Infectious (symbols $\Box$ and O) are shown. In particular, the case fatality rates and detection rates of Table 1 change with the starting point of the prediction, and they determine S, M, and H in the calibration step of section 4.5. This is why the S, M, and H values differ left of the starting points.

The leftmost prediction on day 94 roughly matches the data available up to day 122 in all cases. It has to be taken into account that errors in such models must proliferate exponentially, and then linearly in logarithmic plots. One can see that the Brazil parameters do not change much, while the three predictions for the United States get better. This might be used to assess effectivity of administrative efforts to handle the pandemics.

For an early case in Germany, Figure 10 shows the prediction based on data of day 67, March 27th. The peak of about 35 million hidden and 3.2 million confirmed Infected is predicted on day 121, May 22nd, with about 82,000 casualties at the peak and about 250,000 finally. A good reason to act politically. Note that the real death count is about 8300 on May 23rd, and the prediction of the day, in Figure 8, targets a final count of below 10,000.

Quantitative commitments to predictions are rare in the literature, except for rough estimations of dramatic outbreak scenarios. On April 3rd, after the last public restrictions in Germany of March 22nd, 2020, Germany had 1107 deaths and Khailaie et al. [17] predicted "an order of 10,000 deaths" for the next four weeks. This model predicts 15,500 for May 3rd when run on data of April 3rd, while the true deaths were 6812 on May 3rd, after the interventions worked.

On March 16th, day 54, Ferguson et al. [7] predicted deaths in the order of 250,000 in Great Britain, and 1.1 to 1.2 million in the USA "in the most effective

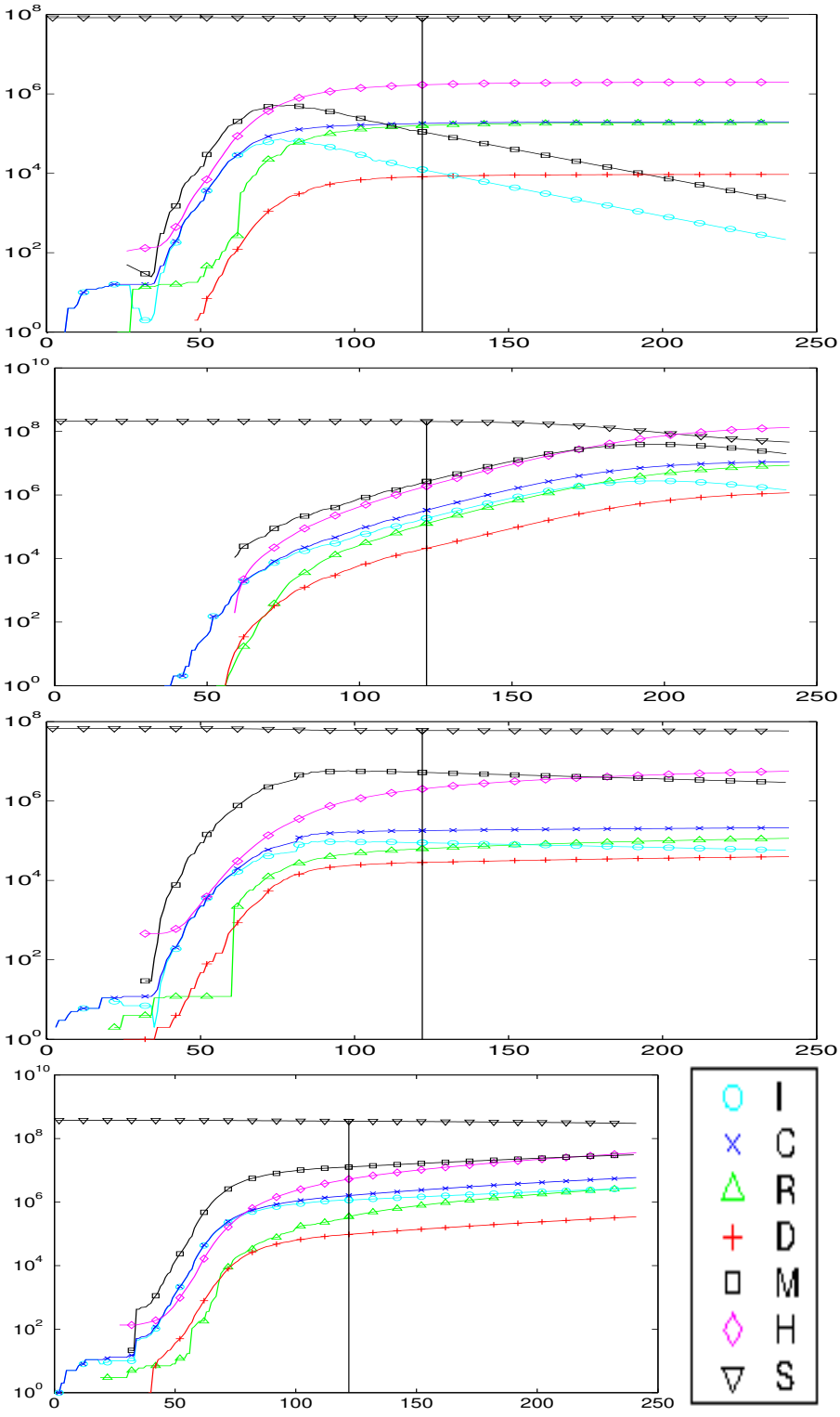

Fig. 8 Predictions for countries Germany, Brazil, France, and US on day 122 marked by the vertical line. The S, M, H values to the left are obtained by calibration, the C, R, D, I values there are the original Johns Hopkins data. Not all data points have marks.

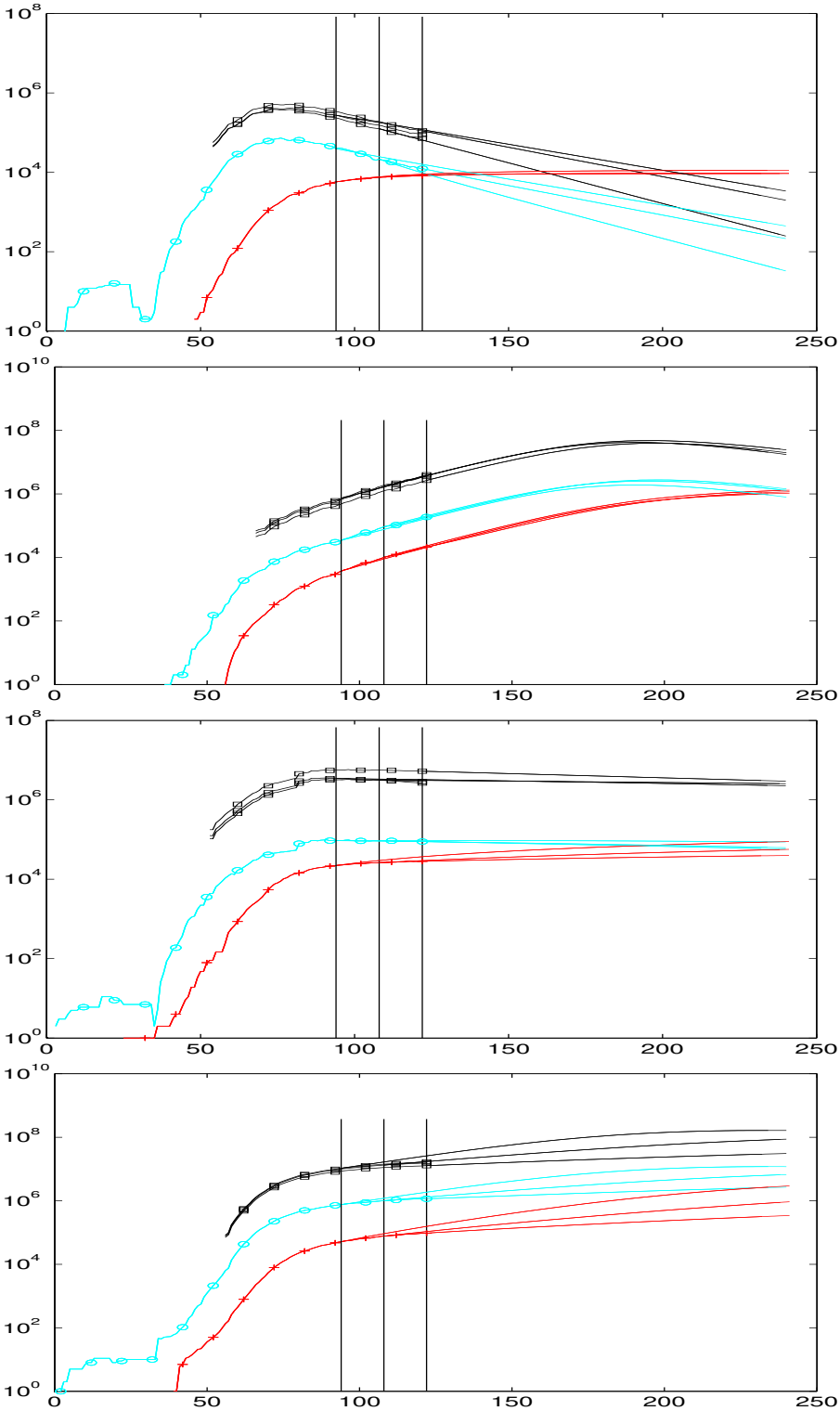

Fig. 9 Predictions for countries Germany, Brazil, France, and USA on days 122, 108, and 94, marked by vertical lines. Legend as in Figure 8, but only M, I, and D shown (M=□,I=O,D=+).

mitigation strategy examined", but not based on the data of the day. In an "unmitigated epidemic" 520,000 deaths in the UK and 2.2 million in the US were predicted, under assumption of $R_0 = 2.4$ and a range of R_0 tested between 2.2 and 2.6. Unfortunately, the model (38) cannot be safely run on day 54 for these countries because there are not enough reliable backlog data. The model can be run if the amount of data used is cut down by choosing $k = 7$ for the k-day rule. Then the predictions on day 54 are more than 30 million deaths for the US and 801,000 for the UK, with a data-based estimation of $R_0 = 6.06$ for the USA and 4.55 for the UK. There is no reasonable data-driven estimate for R_0 that comes close to $R_0 = 2.4$ used by Ferguson et. al. [7] for both countries. They had a much more serious outbreak than assumed by Ferguson et al. on March 16th. See Figure 5 for much later data-based estimates for the US that still are very large.

The use of the Infection Fatality Rate is somewhat different from Streeck et al. [29] and Bommer/Vollmer [2], but results are similar. If the rate 0.0036 of [29] is used in a test run based on data of May 2nd, the estimated number $M_n + C_n$ of total Infected comes out as 1.7 million, while [29] gets 1.8 million by the formula $D_n/0.0036$ for the same day.

The parameter changes by political measures turned out to be rather effective, like in many countries that applied similar strategies. But since parts of the population want to go back to their previous lifestyle, all of this is endangered, and the figures should be monitored carefully.

Of course, all of this makes sense only under the assumption that reality follows the model, in spite of all attempts to design a model that follows reality.

6 Conclusion and Open Problems

So far, the model presented here seems to be useful, combining theory and practically available data. It is data-driven to a very large extent, using only the infection fatality rate from outside for prediction, and the approximation (33) for calibration. On the downside, there is quite a number of shortcomings:

- Like the data themselves, the model needs regular updating. As far as the Johns Hopkins data are concerned, the model updates itself by using the latest data for its internal parameter estimation, but it needs changes as soon as new information on the hidden infections come in.
- There may be better ways of estimating the hidden part of the epidemics. However, it will be easy to adapt the model to other parameter choices. If time series for the unknown variables get available, the model can easily be adapted to being data-driven by the new data.
- The treatment of delays is unsatisfactory. In particular, infected persons get infectious immediately, and the k-day rule is not followed at all places in the model. But the rule is violated as well in the data (Schaback [25]).
- There is no stochastics involved, except for simple things like estimating constants by means, or for certain probabilistic arguments on the side, e.g. in section 4.3.2. But it is not at all clear whether there are enough data to do a proper probabilistic analysis.

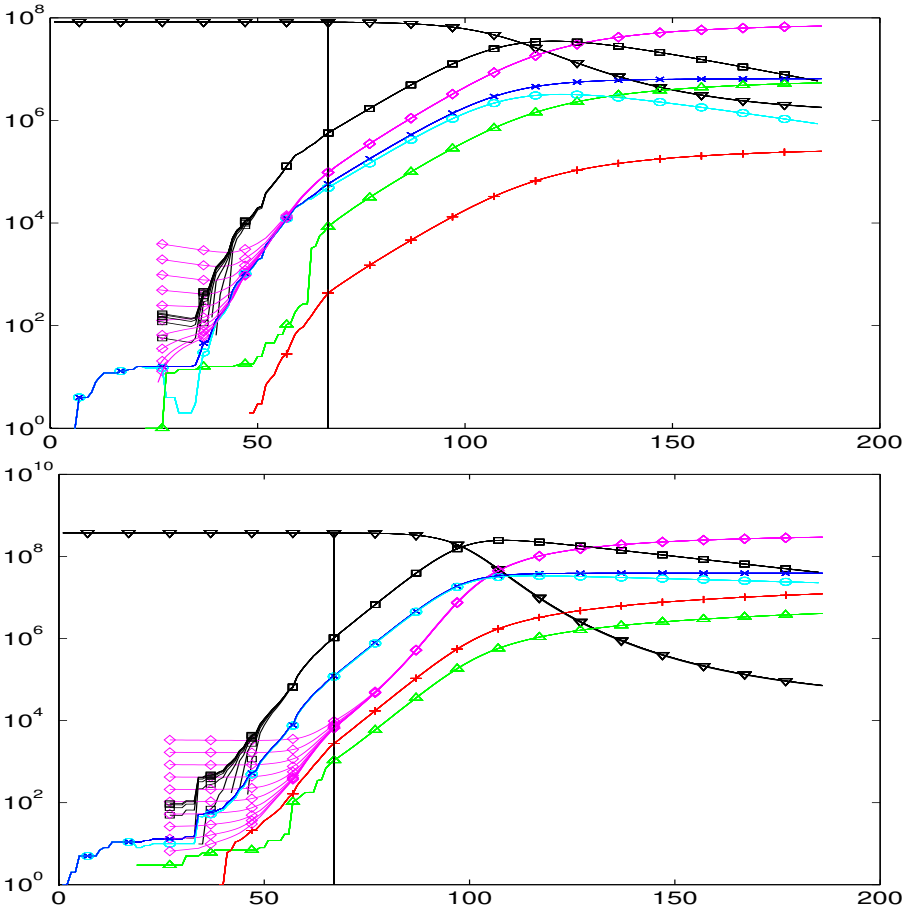

Fig. 10 Predictions for Germany and USA on day 67, March 27th, with varying starting values. Legend as in Figure 8.

- As long as there is no probabilistic analysis, there should be more simulations under perturbations of the data and the parameters. A few were included, e.g. for section 4.3.2 and Figures 7 and 10, but a large number was performed in the background when preparing the paper, making sure that results are stable. However, there are never too many test simulations.
- Totally different models were not considered, e.g. the classical ones with delays (Hoppenstaedt/Waltman [14, 15]), and agent-based approaches (Ferguson et al. [7]) that model infections via contacts and can care for spatial distributions.
- The model needs quite an amount of backward data, making it useless at the very beginning of an outbreak.
- Under certain circumstances, epidemics do not show an exponential outbreak, in particular if they hit only locally and a prepared population. See Figure 11 for the COVID-19 cases in Göttingen and vicinity.

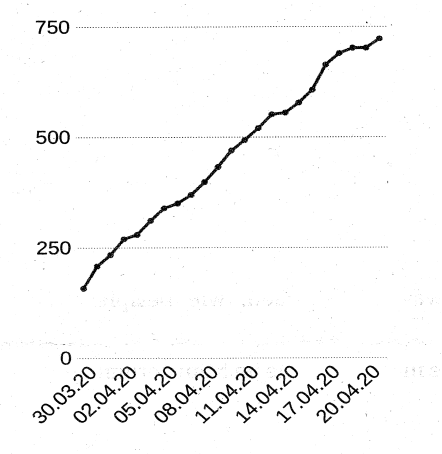

Fig. 11 Infectious in Göttingen city and county, as of April 22nd, 2020 in the local newspaper "Göttinger Tageblatt". No exponential outbreak.

Acknowledgements MATLAB programs and more recent predictions will be on the research website `http://num.math.uni-goettingen.de/schaback/research/group.html` of the author. Special thanks go to Tara Fickle, Reiner Kree, Viola Priesemann, Jalda Schaback, and Wolfgang Warth for various forms of input. All links in the references were verified on June 2nd, 2020.

Conflict of interest

The author declares no conflict of interest.

References

1. Aron, J., Muellbauer, J.: Transatlantic excess mortality comparisons in the pandemic. https://ourworldindata.org/uploads/2020/08/Aron-and-Muellbauer-Transatlantic-excess-mortality-comparison.pdf
2. Bommer, C., Vollmer, S.: Average detection rate of SARS-CoV-2 infections is estimated around six percent (April 2nd, 2020). https://reason.com/wp-content/uploads/2020/04/Bommer-Vollmer-2020-COVID-19-detection-April-2nd.pdf
3. Dandekar, R., Barbastathis, G.: Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. https://www.medrxiv.org/content/10.1101/2020.04.03.20052084v1 (April 6th, 2020). DOI:10.1101/2020.04.03.20052084
4. De Brouwer, E., Raimondi, D., Moreau, Y.: Modeling the COVID-19 outbreaks and the effectiveness of the containment measures adopted across countries (April 19th, 2020). DOI: 10.1101/2020.04.02.20046375
5. Dehning, J., Spitzner, P., Linden, M., Mohr, S., Zierenberg, J., Wibral, M., Wilczek, M., Priesemann, V.: Model-based and model-free characterization of epidemic outbreaks - Technical notes on Dehning et al., Science, 2020 (June 22, 2020)
6. Dehning, J., Zierenberg, J., Spitzner, P., Wibral, M., Pinheiro Neto, J., Wilczek, M., Priesemann, V.: Inferring covid-19 spreading rates and potential change points for case number forecasts (May 4th, 2020). ArXiv:2004.01105v2

7. Ferguson, N., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., van Elsland, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P., Walters, C., Winskill, P., Whittaker, C., Donnelly, C., Riley, S., Ghani, A.: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. DOI: 10.25561/77482. Imperial College London (16-03-2020)
8. Friston, K.J., Parr, T., Zeidman, P., Razi, A., Flandin, G., Daunizeau, J., Hulme, O.J., Billig, A.J., Litvak, V., Moran, R.J., Price, C.J., Lambert, C.: Dynamic causal modelling of covid-19 (April 9th, 2020)
9. Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., Colaneri, M.: Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. https://doi.org/10.1038/s41591-020-0883-7 (April 22nd, 2020). www.nature.com/naturemedicine
10. GitHub: Covid-19 repository at github. https://github.com/CSSEGISandData/COVID-19 /tree/master/csse_covid_19_data/csse_covid_19_time_series (2020)
11. An der Heiden, M., Buchholz, U.: Modellierung von Beispielszenarien der SARS-CoV-2-Epidemie 2020 in Deutschland. Bekanntmachungen des Robert Koch-Instituts (March 3rd, 2020). DOI: 10.25646/6571.2, https://edoc.rki.de/handle/176904/6547.2
12. An der Heiden, M., Hamouda, O.: Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland - Nowcasting. Epidemiologisches Bulletin **17**, 10–15 (2020)
13. Höhle, M., an der Heiden, M.: Bayesian nowcasting during the STEC 0104:H4 outbreak in Germany, 2011. Biometrics pp. 993–1002 (2014)
14. Hoppensteadt, F., Waltman, P.: A problem in the theory of epidemics I. Math. Biosci. **9**, 71–91 (1970)
15. Hoppensteadt, F., Waltman, P.: A problem in the theory of epidemics II. Math. Biosci. **12**, 133–145 (1971)
16. Kermack, W., McKendrick, A.: A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society A **115**, 700–721 (1927)
17. Khailaie, S., Mitra, T., Bandyopadhyay, A., Schips, M., Mascheroni, P., Vanella, P., Lange, B., Binder, S., Meyer-Hermann, M.: Estimate of the development of the epidemic reproduction number r_t from Coronavirus SARS-CoV-2 case data and implications for political measures based on prognostics (April 7th, 2020). DOI: 10.1101/2020.04.04.20053637, https://www.medrxiv.org/content/10.1101/2020.04.04.20053637v1
18. Kucharski, A., Russell, T., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R.: Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis. **20**, 553–558 (May 1st, 2020). DOI: 10.1016/S1473-3099(20)30144-4
19. Maier, B.F., Brockmann, D.: Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. Science **368**(6492), 742–746 (April 2020). DOI 10.1126/science. abb4557. URL `http://dx.doi.org/10.1126/science.abb4557`
20. Mohring, J., Wegener, R., Gramsch, S., Schöbel, A.: Prognosemodelle für die Corona-Pandemie (April 29th, 2020). Fraunhofer-Institut für Techno- und Wirtschaftsmathematik ITWM Kaiserslautern, https://www.itwm.fraunhofer.de/content/dam/itwm/de/documents/PressemitteilungenPDF/2020/20200429_Bericht_Prognosemodelle-für-die-Coronapandemie.pdf
21. Phipps, S., Grafton, R., Kompas, T.: Estimating the true (population) infection rate for COVID-19: A backcasting approach with Monte Carlo methods. https://doi.org/10.1101/2020.05.12.20098889
22. Robert-Koch-Institut: Erläuterung der Schätzung der zeitlich variierenden Reproduktionszahl r. Tech. rep., Robert-Koch-Institut (2020). 15.05.2020, https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/R-Wert-Erlaeuterung.html
23. Robert-Koch-Institut: SARS-CoV-2 Steckbrief zur Coronavirus-Krankheit-2019 (COVID-19). Tech. rep., Robert-Koch-Institut (2020). 22.05.2020, https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html
24. Schaback, R.: On COVID-19 Modelling. Jahresberichte der Deutschen Mathematiker-Vereinigung **122**, 167–205 (2020). June 11th, 2020
25. Schaback, R.: Modelling recovered cases and death probabilities for the COVID-19 outbreak (March 26th, 2020). URL `https://arxiv.org/abs/2003.12068`
26. Sentker, A..: Bloß raus hier! DIE ZEIT (2020). April 16th, 2020, page 20
27. Silverman, J., Hupert, N., Washburne, A.: Using influenza surveillance networks to estimate state-specific prevalence of SARS-CoV-2 in the United States. Science Translational Medicine **12**, eabc1126 (July 29th 2020). DOI 10.1126/scitranslmed.abc1126

28. Stafford, N.: Covid-19: Why Germany's case fatality rate seems so low. BMJ p. 369:m1395 (April 7th, 2020). DOI 10.1136/bmj.m1395
29. Streeck, H., Schulte, B., Kuemmerer, B., Richter, E., Hoeller, T., Fuhrmann, C., Bartok, E., Dolscheid, R., Berger, M., Wessendorf, L., Eschbach-Bludau, M., Kellings, A., Schwaiger, A., Coenen, M., Hoffmann, P., Noethen, M., Eis-Huebinger, A.M., Exner, M., Schmithausen, R., Schmid, M., Hartmann, G.: Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event (May 8th, 2020). DOI 10.1101/2020.05.04.20090076
30. The Economist: Tracking covid-19 excess deaths across countries (2020)
31. Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P.G.T., Fu, H., Dighe, A., Griffin, J.T., Baguelin, M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunubá, Z., FitzJohn, R., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Laydon, D., Nedjati-Gilani, G., Riley, S., van Elsland, S., Volz, E., Wang, H., Wang, Y., Xi, X., Donnelly, C.A., Ghani, A.C., Ferguson, N.M.: Estimates of the severity of coronavirus disease 2019: a model-based analysis (June 1st, 2020). URL `https://doi.org/10.1016/S1473-3099(20)30243-7`
32. Viglione, G.: How many people has the coronavirus killed? Nature **585**, 22–24 (Sep. 1st, 2020)
33. WHO: Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). https://www.who.int/publications-detail/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19) (Feb. 28th, 2020)
34. Wikipedia: Compartmental models in epidemiology. https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology (June 2nd, 2020)

Additions

This is a very informal appendix containing additions, remarks, and changes to the original version of the paper [24] for the DMV. This version is of Jan. 15th, 2021 . No further versions are planned, because the models need too many additions that require plenty of additional data and take too much of my time:

1. including increased vaccination,
2. including increased testing,
3. incorporating realistic delays, and
4. modelling contacts in a much more detailed way, including mobility, age groups, and schooling.

Earlier versions of these additions are on my webpage[2] together with MATLAB programs. Readers only interested in new predictions should directly go to section 7, if they are familiar with the notation in the basic article.

Since the various additions turned out to be unwieldy, they needed some rearrangement. They are now grouped as follows:

- For the impatient: section 7 *New predictions* provides new examples based on the latest Johns Hopkins data, incorporating the changes of models and algorithms described in later sections.
- Mathematical additions (section 8):
  - *Additional Explanations* (section 8.1) provides further information about certain parts of the original paper:
    - *on the asymptotic exponential decay* (section 8.1.1),
    - *on the peak location* (section 8.1.2), and
    - *on global existence of solutions* (section 8.1.3).
  - *Constant new Confirmations* (section 8.2) deals with the situation of roughly constant or slowly increasing new infections, as is the case for various countries from mid-July 2020. This is a modification of the Johns Hopkins Data Model. It predicts that a roughly constant rate of new infections implies a roughly constant rate of new COVID-19 deaths.
  - *Shortcomings of the models* (section 8.3) describes problems that arose after publication of the basic text. It mainly concerns two things:
    - instabilities due to low numbers of new casualties behind peaks, or due to data flaws like non-monotonicity of data for cumulative variables,
    - replacing constants by time series.
  - *Imputation and Nowcasting* (section8.4) mimics a technique used by the Robert Koch Institute (RKI) to improve the data quality. It helps to overcome the aforementioned shortcomings.
  - The *validity of the k-day rule* is checked in section 8.5 for various countries.
- *Non-mathematical additions to the text* (section 9) includes new references and comments, e.g. honouring the improvement in the way the media deal with mathematical concepts, or comments on recent literature.

[2] http://num.math.uni-goettingen.de/schaback/research/group.html

– *Changes in programming* (section 10) describes why certain technical modifications in the algorithms were necessary. Changes based on mathematics are in sections 8.3 and 8.4.

7 New Predictions

7.1 Latest Version of January 15th, 2021

Again, we only add new plots, and do not change most of the text from the earlier versions following below. Johns Hopkins data were used up to day 359 (Jan. 15th, 2021), but since the smoothing process decribed in section 8.4.2 moves cases backwards for up to seven days, the smoothed data were used only from day 352 on. Since the United States stopped to deliver values for the Recovered from day 328 on, the Recovered had to be estimated by applying the 14-day rule, as for France, Spain, and the United Kingdom.

Figure 12 extends Figure 14 of November 2020 and shows how countries struggle to get the derivatives at the upper margin close to one, aiming at $R_0 = 1$. Since tick marks are by weeks, counting back from the top, one can see how they tried to improve since mid-December 2020. One might trace changes of the derivatives back to lockdown measures taken.

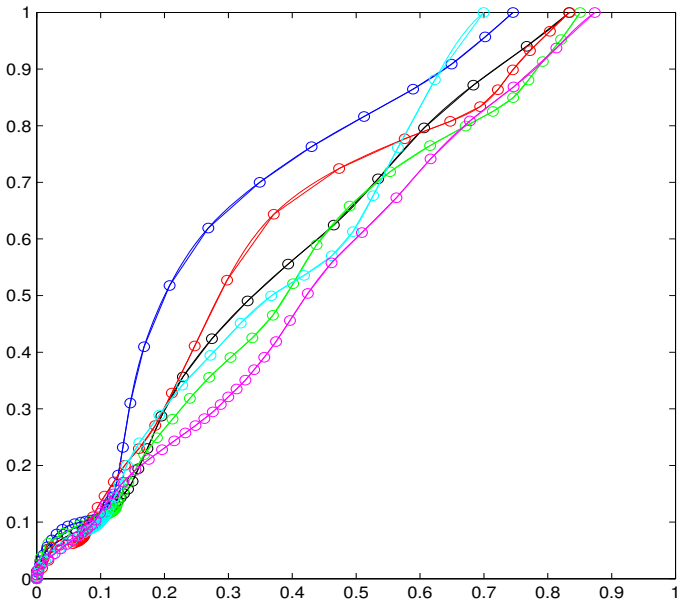

Fig. 12 C as a function of $R+D$ up to day 352 (Jan 8th), normalized. Countries are Germany (black), Italy (blue), Spain (green), France (red), UK (cyan), USA (magenta). Tick marks are for weeks, counting backwards from the top margin. The derivative is an estimate of R_0.

The actual situation for Germany is in Figure 13, Though the Infectious will not have a new peak, they will decrease rather slowly, leading to an expected total death toll of about 110,000 in March 2020, if there is no counteraction. This is realistic, the current total casualities being at about 45,000 with a daily increase of about 1000, as of mid-January. Compare with the fast decrease after the first peak, caused by a complete lockdown. The measures taken in December are by far not satisfactory to get a similar decline of the Infectious.

Similar plots for other countries give expected total casualties 490000 for the United States, 125000 for the United Kingdom, 100000 for France, and 72000 for Spain.

But these results are rather unreliable, because they ignore increasing vaccination and increased testing.

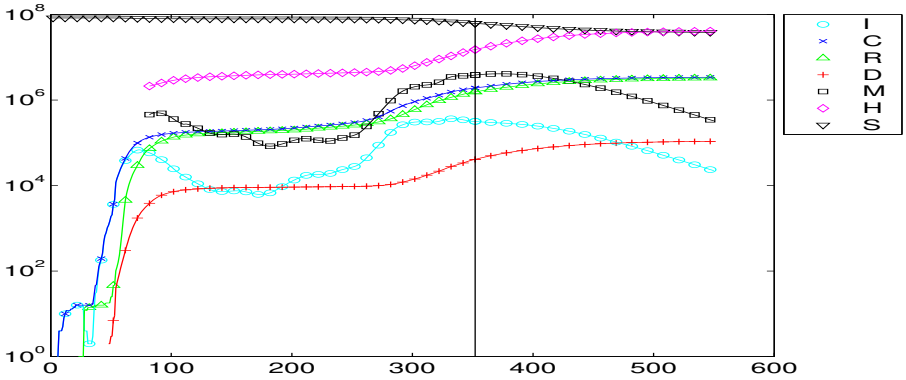

Fig. 13 Prediction for Germany from day 352.

7.2 Version of November 24th, 2020

Figure 14 extends Figures 20 and 16 to show the slopes describing R_0 from Johns Hopkins data. The situations for Germany and Italy have improved somewhat, in contrast to the others. Here and for the other new plots, Johns Hopkins data were used up to day 307 (Nov. 24th), but since the smoothing process decribed in section 8.4.2 moves cases backwards for up to seven days, the smoothed data were used only from day 300 on.

Figure 15 shows the current peak prediction for Germany by the full model. The infection peak is predicted in mid-February with about 875000 registered Infectious, which is still intolerable. Intermediate test runs were even worse, but the new political actions showed some improvement. Still, the administrative actions of early November seem to be insufficient. There is some hope that the mid-December data show some improvement, because the hidden part of the model needs some time to react to new data. The final death toll may rise slowly to about 90000 well after the second

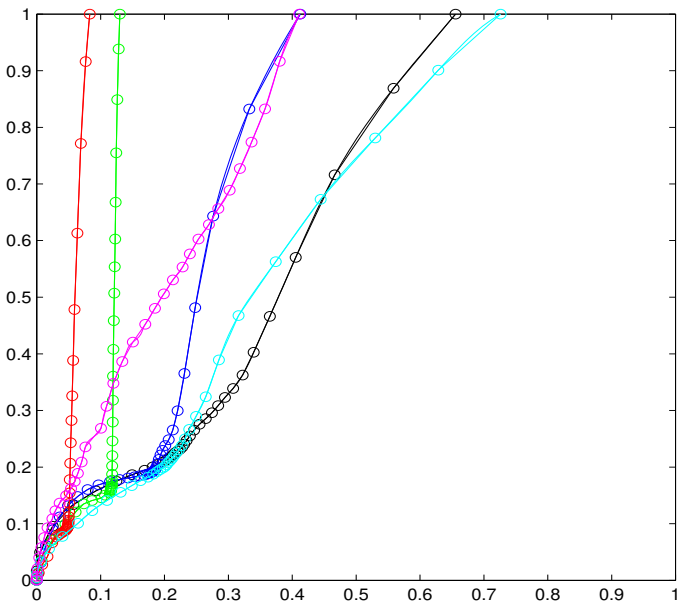

Fig. 14 C as a function of $R+D$ up to day 300 (Nov 17th), normalized. Countries are Germany (black), Italy (blue), Spain (green), France (red), UK (cyan), USA (magenta). Tick marks are for weeks. The derivative is an estimate of R_0.

peak, but this slow increase may be tolerable by the health system. Note that the RKI reported about 25000 casualties due to influenza in Germany in 2017/18, a very bad winter. Anyway, all of this is too serious to delay counteractions.

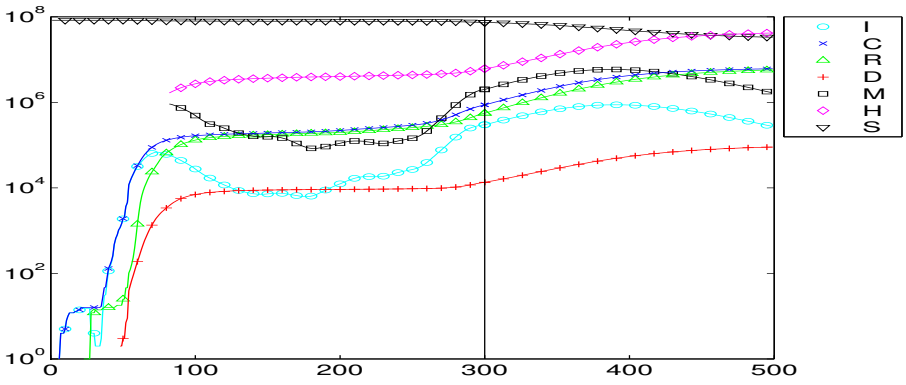

Fig. 15 Prediction for Germany from day 300.

7.3 Version of October 16th, 2020

Here, we only add new plots, because the situation for several countries has changed considerably since mid-September. We just add new versions of the figures in the September version, and comment them. The start is made by a new version of Figure 20. The situation of Germany and Italy got worse, the UK have improved somewhat, while USA, Spain and France stay in the same scary situation. Recall that curves should have small derivatives, in particular smaller than one, since the derivative is connected to the Reproduction Number. The UK results are questionable, because the Johns Hopkins data have nothing on the Recovered, and the data for France and Spain have flaws that are hard to come over, see remarks in earlier updates of the paper.

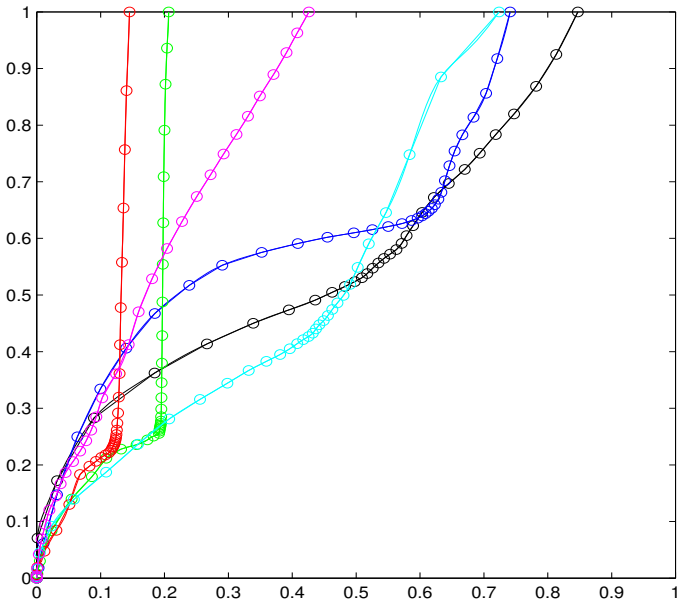

Fig. 16 C as a function of $R+D$ up to day 268 (Oct 16th), normalized. Countries are Germany (black), Italy (blue), Spain (green), France (red), UK (cyan), USA (magenta). Tick marks are for weeks. The derivative is an estimate of R_0.

Figure 21 is replaced by Figure 17. This shows the serious change in early October.

The same holds for Table 2, summarizing full model test runs, replacing Table 3. As already mentioned in earlier updates, the data for France and Spain have plenty of flaws that still make predictions very unstable, but these are disastrous anyway. The model is quite sensitive to recent changes, and this can be seen for the dramatic negative changes for Germany and the positive changes for Denmark. Compare with Figures 24 and 25. The death toll for Germany jumped from about 10500

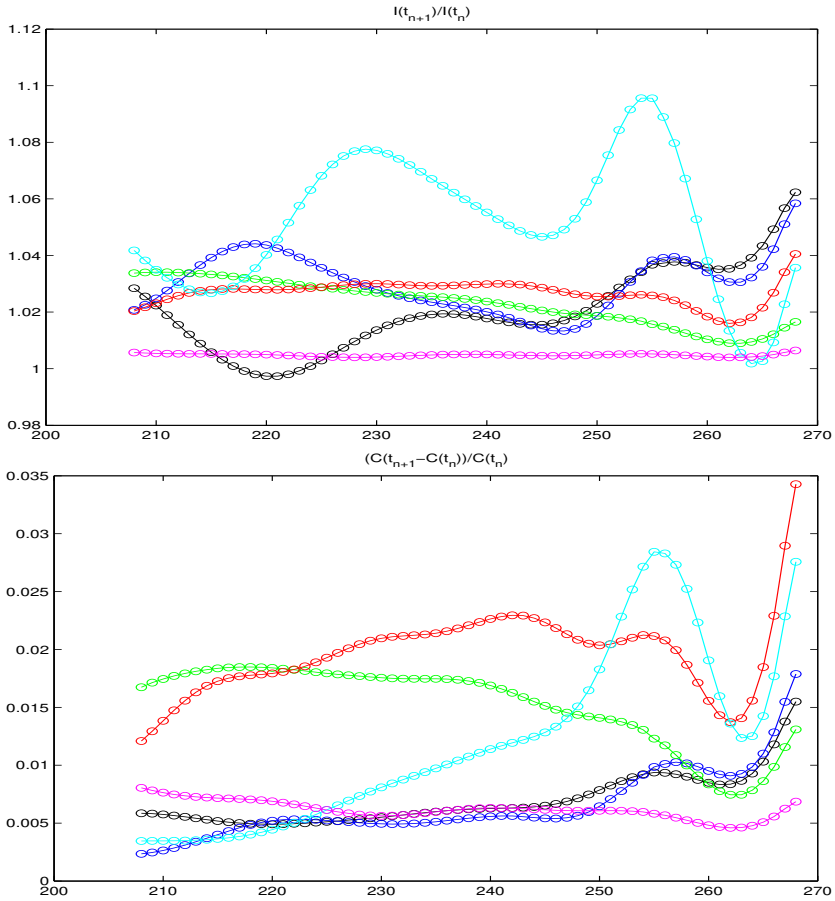

Fig. 17 Daily multipliers for Infectious and new Confirmed, now up to day 268, October 16th. Countries are Germany (black), Italy (blue), Spain (green), France (red), UK (cyan), USA (magenta).

to over 110000, while the expected casualties for Denmark drop from about 5800 to 1640. Looking at the changes of R estimates, and keeping in mind that exponential behaviour must be expected, this is not really surprising. Germany's new measures taken in mid-October might not be drastic enough to prevent an exponential outbreak with $R \approx 2$. The United States roughly stay on their path towards about a million COVID-19 casualties.

Country	R_{full}	R_{JH}	R_{SCH}	δ_{start}	δ_{final}	max(D)	max(I)	attained at day
Germany	2.09	2.18	1.80	0.05	0.15	117294	2279771	13-Feb-2021
Italy	3.51	4.00	2.16	0.01	0.12	185297	1420433	09-Apr-2021
Spain	10.16	23.76	5.94	0.02	0.12	465019	3262681	23-Jan-2022
France	6.72	29.43	3.93	0.01	0.15	894326	4555292	09-Jan-2022
UK	2.15	1.30	1.51	0.01	0.14	66192	289804	23-Dec-2020
US	2.23	2.04	1.39	0.02	0.10	962632	6729141	16-May-2021
Denmark	1.15	1.06	0.90	0.03	0.17	1640	6770	12-Mar-2021

Table 2 Various R_0 estimates, peak predictions, and maximal counts of deaths (D) and Infectious (I), using data up to day 268.

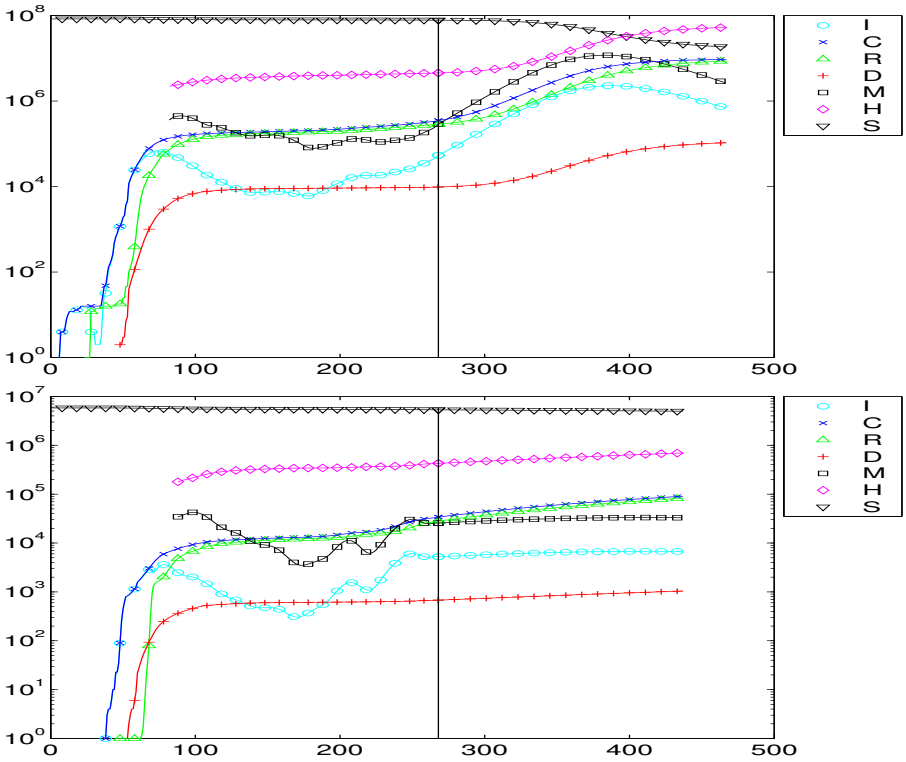

Fig. 18 Predictions for Germany and Denmark from day 268.

7.4 Version of September 2020

7.4.1 Data Quality

The data quality has degraded even more since July. This is fatal for any model that tries to stay as close as possible to the given data. The Johns Hopkins data of Spain have no new Recovered since day 118 (May 19th) up to day 230 (Sep. 11th, as of this

writing), no new deaths between days 138 and 149 and the cumulative Confirmed drop from 213024 to 202990 between days 93 and 94, while the cumulative fatalities drop from 28752 to 26834 between days 124 and 125. These data flaws were too serious to estimate the data for the Recovered under any reasonable k-day rule, using the older versions of the techniques of sections 4.3.2 and 8.4.2. See Figure 19 for recent results concerning France and the United Kingdom, using the improved software of section 8.4.2 for data correction and smoothing.

Many countries now have unexpectedly low death rates, which is good news, but causes problems to data-driven models that rely on well-reported casualties. As opposed to the peak situation, the D and R curves (red and green) do not follow the I curve (black) for the Infectious with a reasonable delay, e.g. in Figure 19 for France. This makes fatality and recovery data somewhat questionable, but it is possible that new Confirmed are now much younger, endangered subpopulations are better protected, and health systems are not under stress. The predictions below will get worse as soon as fatality rates go up again. We shall deal with the issue of very low death rates for high infection rates at various places below. As some form of counteraction, recent data on excess mortality (see section 9.2, Table 7) have been implemented as factors on the Johns Hopkins figures for casualties. But, so far, they were not entered as time series.

7.4.2 Estimates of R_0 by the Johns Hopkins Data Model

Because of the 0/0 instability of (24) for extremely low numbers of new deaths and recoveries behind peaks, Figure 5 will not be updated.

After quite some experimentation, a better solution was based on the observation that (24) is a coarse way to estimate the derivative of the Confirmed C as a function of the Removed $R+D$, consisting of Dead D and Recovered R. If the pre-processing technique of section 8.4.3 is applied, this function behaves reasonably well, see Figure 20 for Germany, Italy, France, and USA. Since both C and $R+D$ are monotonic over time, the derivative of the curves shows the behaviour of the R_0 estimate over time, the time scales marked on the curves being different for each country. The C and $R+D$ values for each country were normalized per country by dividing through their common maximum, to keep the derivative unchanged. Tick marks for each week show the time behaviour, backwards from the top. The derivative at the top line is the up-to-date approximation of R_0. The Infectious $I = C-R-D$ are the length of the vertical line between a C value and the diagonal. Countries with few Infectious keep close to the diagonal. Curves further away have an overweight of C over $R+D$ for whatever reason, e.g. more testing or bad data transmission for Recovered. The derivatives of $C(R+D)$ at the upper endpoints as estimates of R_0 by (24) for day 236 (September 14th, 2020) come out to be 1.22 for Germany, 2.28 for Italy, 2.05 for the United Kingdom, and 2.09 for USA by a least-squares backward fit on the previous week by a linear polynomial.

The near-to-vertical curves for France and Spain deserve a closer look. For the last five weeks before day 236, France had hardly any new Deaths or Recovered while the Confirmed nearly doubled. This is hard to believe and in conflict with any reasonable k-day rule, unless there is a serious change in the testing strategy and/or the infected

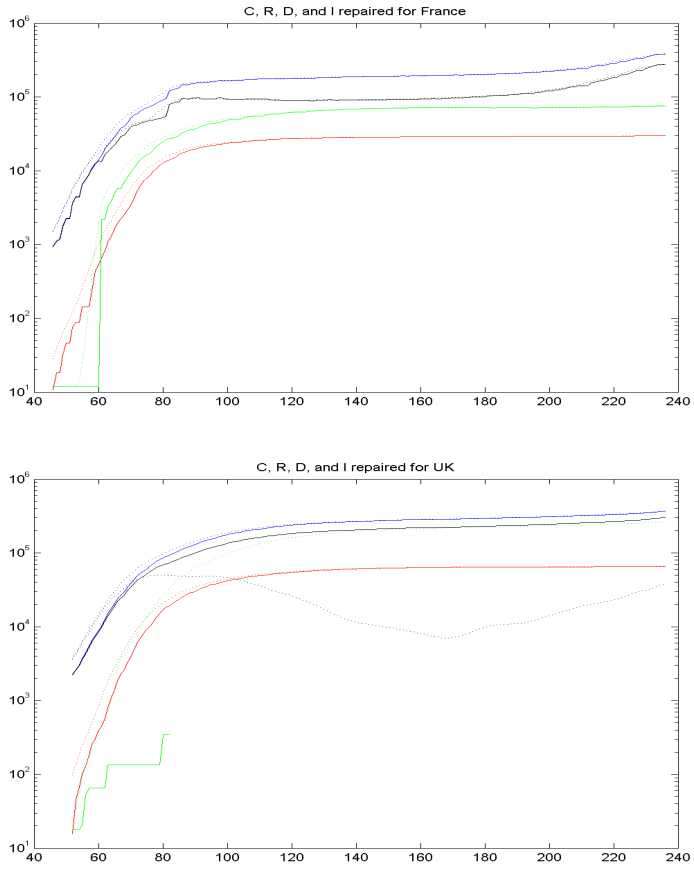

Fig. 19 The Johns Hopkins data for France and the United Kingdom, with repair, up to day 236 (September 14th). Dotted: Repaired data. Compare to Figure 38 for the July version of France.

subpopulation. The function $C(R+D)$ jumps up for this reason. The situation for Spain is similar.

In general, one can see the time regions (except for USA) where the countries kept the derivative close to one, but only Germany is currently better off. This sort of figure replaces Figure 5 of the basic paper. It shows how countries march through the pandemic, and it will be interesting to see how these curves go on in the future.

The instability effect for R_0 is connected to extremely low estimates via (23) for γ in $R_0 = \frac{\beta}{\gamma}$, leading to the seemingly paradoxical situation that R_0 shoots up while fatality rates go down. If one goes for $\beta - \gamma$ instead, as in (11), one should plot the quotients I_{n+1}/I_n to see the increase of the Infectious. The media usually monitor $(C_{n+1} - C_n)/C_n$, and Figure 21 shows these two time series between days 176 and 236 (July 16th to September 14th). Both have no direct connection to the

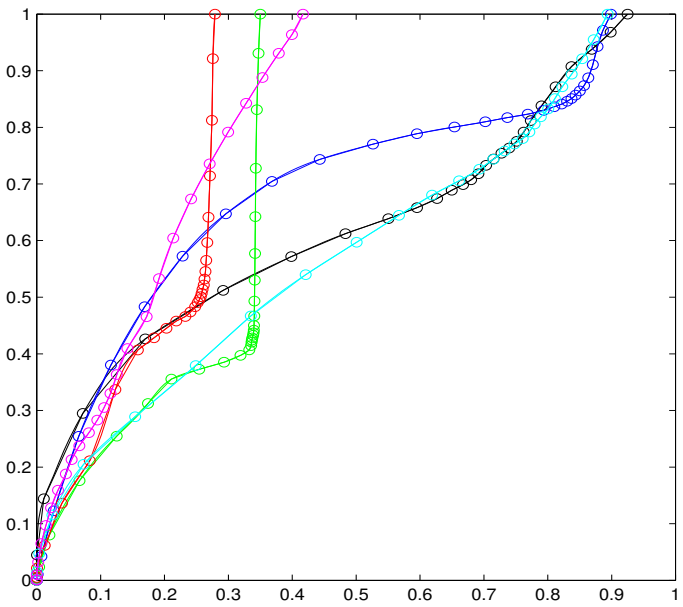

Fig. 20 C as a function of $R+D$ up to day 236 (Sep 14th), normalized. Countries are Germany (black), Italy (blue), Spain (green), France (red), UK (cyan), USA (magenta). Tick marks are for weeks. The derivative is an estimate of R_0.

Basic Reproduction Number, but they describe the actual situation quite well. Raw Johns Hopkins data for $(C_{n+1}-C_n)/C_n$ give Figure 22, showing the dramatic effect of preprocessing following section 8.4.3. The awful flaws in data for France and Spain show up dramatically, polluting (23).

7.4.3 Full Model Predictions

The previous section showed that the R_0 predictions from the Johns Hopkins data worked better when taking a smoothed derivative of $C(R+D)$ instead of the crude discretization (24). The same can be tried for the derivative of $-S(C+H)$ of the full model, following (36). But the model equations (38) need β and γ separately, leading to a different value via β/γ. We shall give numerical examples later, together with the predictions.

There were some serious changes to the full model for the September version, and these are described in detail in section 8.3.4 and 8.3.5. Their motivation was the rapid change of observed case fatalities over time, forcing to treat fatality rates as time series, not constants. As a consequence, detection rates were not constant anymore. These changes were applied directly to the calibration phase of the full model, leading indirectly to different starting values when making predictions.

During the calibration phase, the R_0 estimates by both models can be compared. Except for France and Spain with the aforementioned flaws in Johns Hopkins data, the two estimates match reasonably well, see Figure 23 for Germany (left) and the

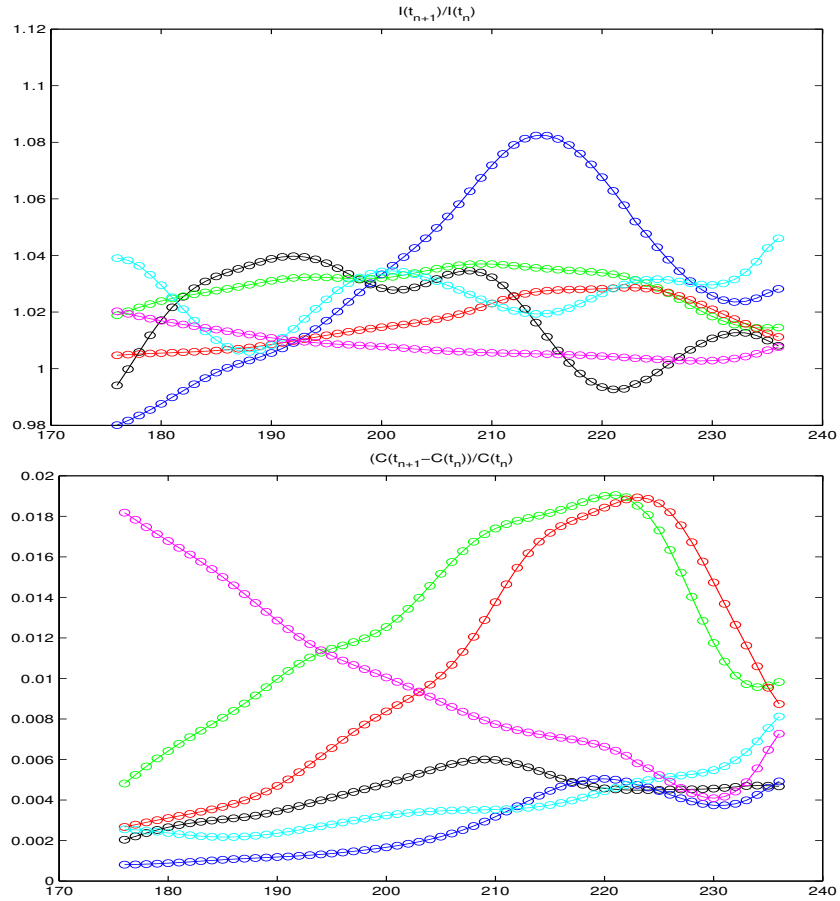

Fig. 21 Daily multipliers for Infectious and new Confirmed. Countries are Germany (black), Italy (blue), Spain (green), France (red), UK (cyan), USA (magenta).

United States (right). Table 3 contains estimates R_{full} for the full model (as quotient β/γ), of R_{JH} for the Johns Hopkins data model, and R_{SCH} as the final derivative of $-S(C+H)$ at the end of calibration, i.e. at day 236. The detection rates at the beginning and at the end of the calibration phase are δ_{start} and δ_{final}. Note, however, that the formula (43) is implemented with $\delta = 0.2$, putting an upper bound of 20% to the detection rate when case fatality rates go down. The starting infection fatality rate γ_{IF} in (43) was set to 0.005 like in all earlier simulations. The peak in Table 3 is given by the day of maximal Infectious I in the final column, while the maximal number of deaths in the penultimate column always occurs at the end of a (suppressed) simulation run much longer than the peak. Of course, all entries will vary as soon as the data change or parameters like in (43) are varied. Many test runs were performed, and the examples presented here are typical. Desastrous cases stayed desastrous even under

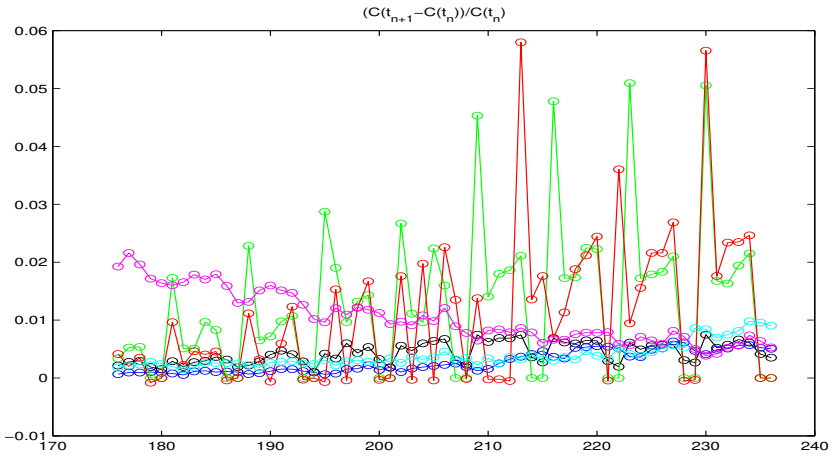

Fig. 22 Same as the lower plot in Figure 21, but for raw Johns Hopkins data.

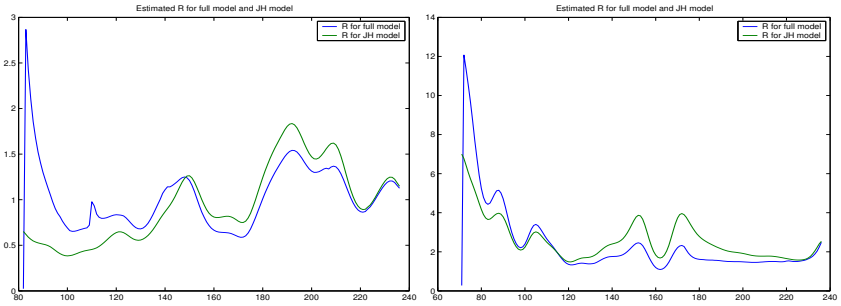

Fig. 23 R_0 estimates by the Johns Hopkins data model (green) and the full SIR model (blue), for Germany and USA, during the calibration phase.

variations of parameters. Nearly all cases show that COVID-19 will be a long-term challenge until vaccination is available.

The rows of Table 3 for France and Spain show the instability of R_{JH} due to missing data for Recovered R and deaths D over longer time periods. The Recovered for the United Kingdom had to be estimated anyway by the 14-day rule.

For illustration, we present the full predictions in Figures 24 and 25, with the same legends as in the paper for Figure 8. Countries behind their peak, with R_0 close to one and low fatality rates are heading for a very long presence of COVID-19 cases at low death rates, as also motivated by section 8.2.

There are also countries that currently are heading for a second peak that is more serious than the first. See Figure 25 for Denmark. The same holds for France and Spain, but there the data have serious flaws when estimating the R_0 values from the Johns Hopkins data. There, small $D+R$ values spoil the applicability of the Johns Hopkins Data model. The full model has larger values $R+H$ for the Recovered, and R_0 estimates get more reasonable, as shown in Table 3. To see how the full model

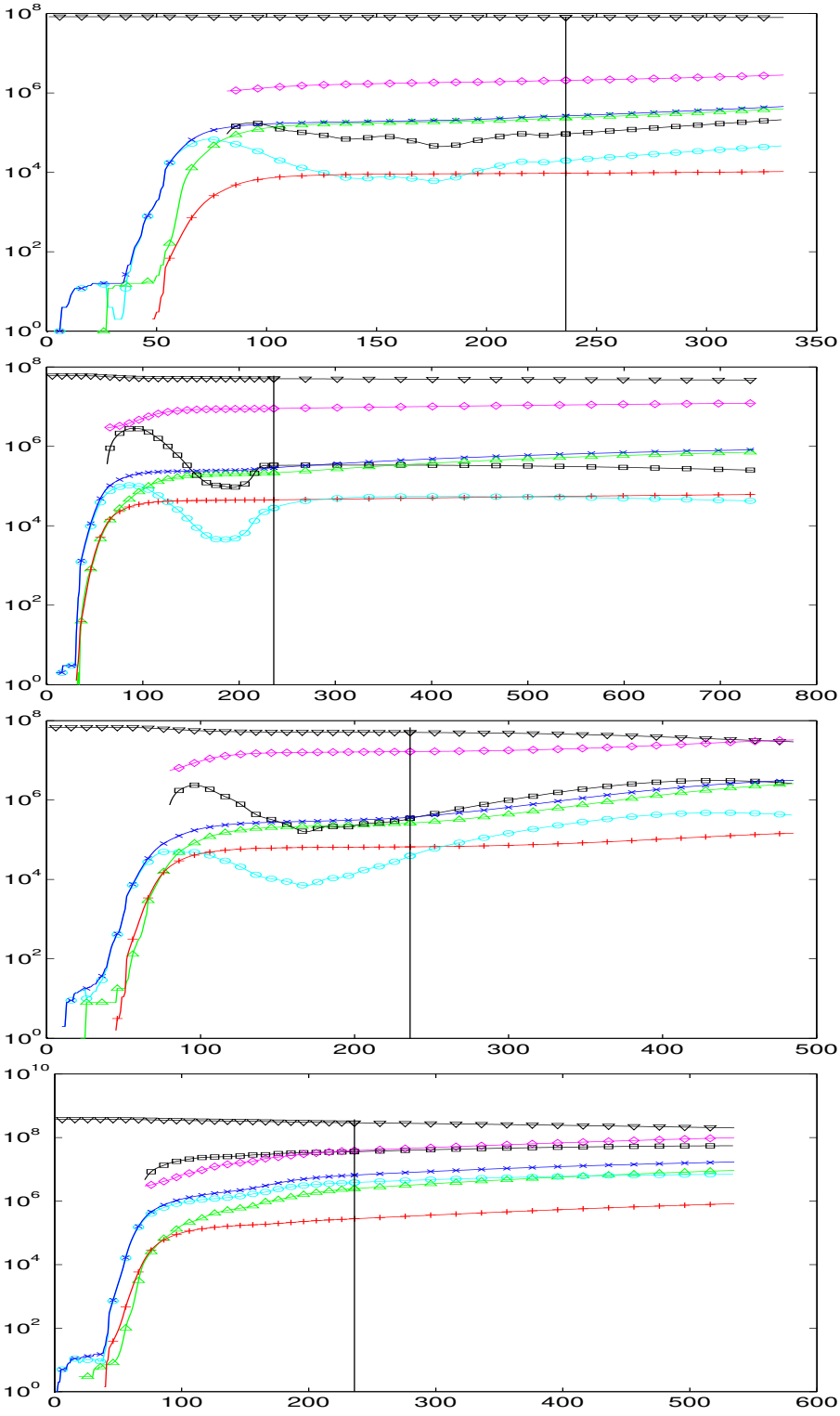

Fig. 24 Predictions for Germany, Italy, UK, and USA from day 236.

Country	R_{full}	R_{JH}	R_{SCH}	δ_{start}	δ_{final}	max(D)	max(I)	attained at day
Germany	1.17	1.21	1.07	0.10	0.18	10496	46832	23-Dec-2020
Italy	1.33	2.34	1.33	0.03	0.14	72661	113059	13-Dec-2021
Spain	6.68	27.12	4.98	0.03	0.16	1177626	4644547	26-Mar-2022
France	3.99	27.17	3.89	0.02	0.18	1025947	5890762	21-Oct-2023
UK	1.94	2.03	1.36	0.02	0.13	146808	476752	04-Apr-2021
US	1.94	2.04	1.33	0.04	0.12	827132	6963843	11-Jul-2021
Denmark	1.96	2.28	1.82	0.09	0.19	5809	197332	21-Dec-2020

Table 3 Various R_0 estimates, peak predictions, and maximal counts of deaths (D) and Infectious (I).

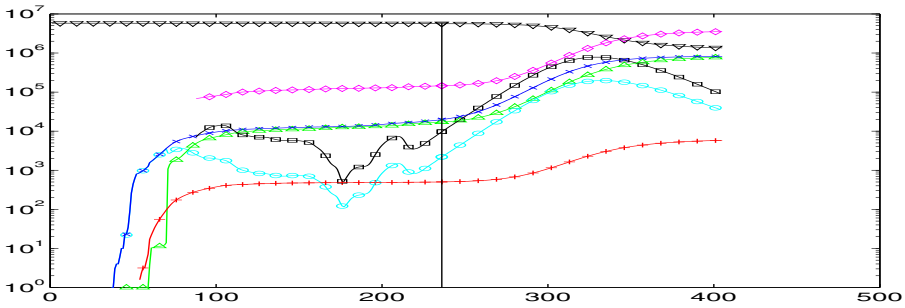

Fig. 25 Predictions for Denmark from day 236.

handles these countries, Figure 26 shows both the long-term prediction and a short-term prediction of 100 days ahead. Both countries are in a state that is worse than the first peak. This modelling used unreliably low numbers of deaths and recoveries given in the Johns Hopkins data set, but predictions would be even more scary if the reported case fatalities were replaced by more plausible ones.

Summarizing, countries have not succeeded in keeping infections at bay. A small and slow increase of daily new infections seems to be tolerated, leading to a small and slow increase of daily new casualties by section 8.2 and the following paragraph. The full SIR model then stays close to an $R_0 \approx 1$ situation.

7.4.4 Predictions for Constant New Confirmed

Section 8.2 has a variation of a Johns Hopkins Data Model that makes predictions in case of a roughly constant number of new Confirmed. In view of the situation in September, we now assume 2500 new daily Confirmed for Germany, and take recent estimates for case fatality and case recovery. Then the Infectious will level out in about a month, and there will be 8 new COVID-19 casualties per day, nationwide. We add Figures 27 and 28 replacing Figures 34 and 35 of the July version below, but now with better preprocessing.

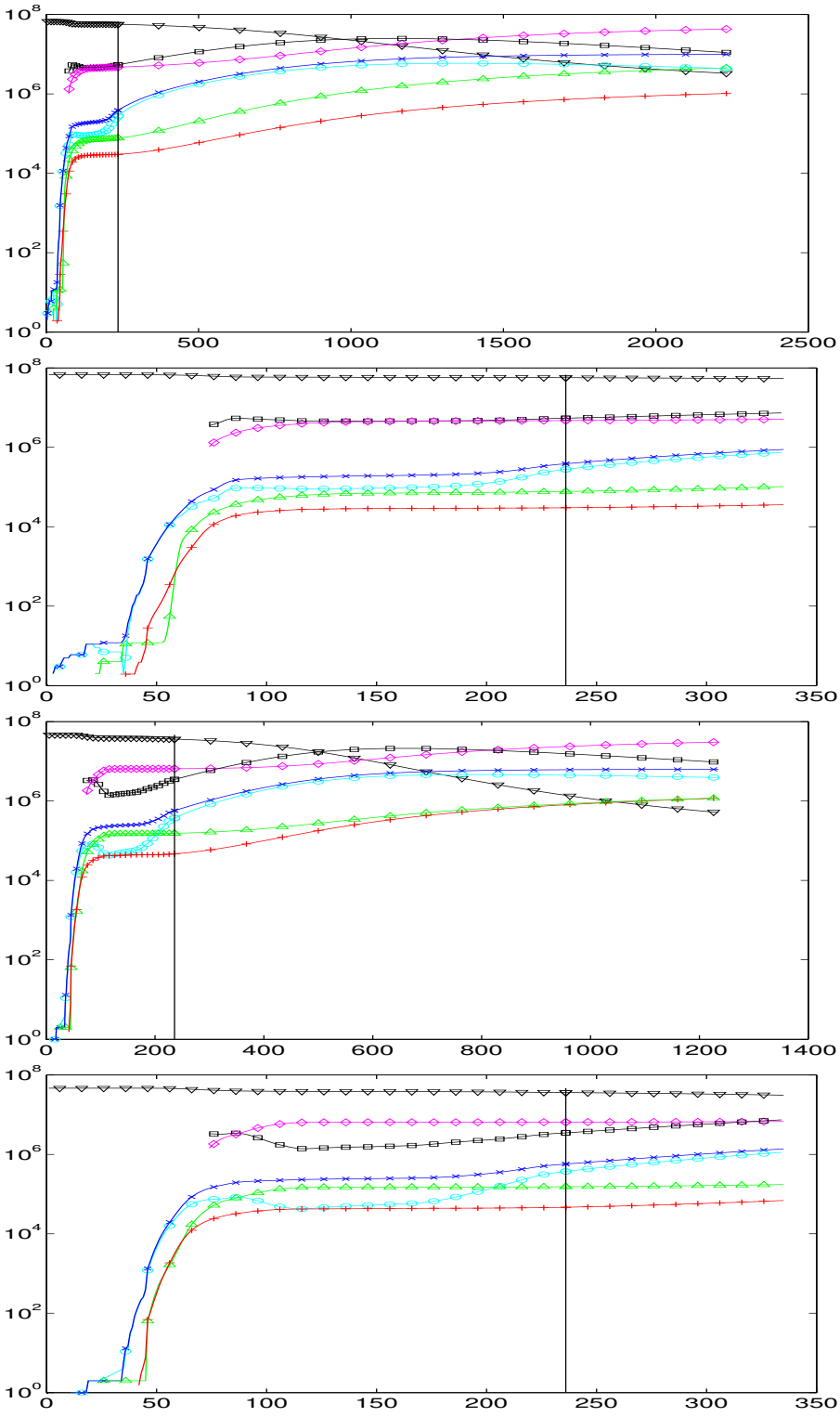

Fig. 26 Predictions for France and Spain, from day 236, long- and short-term.

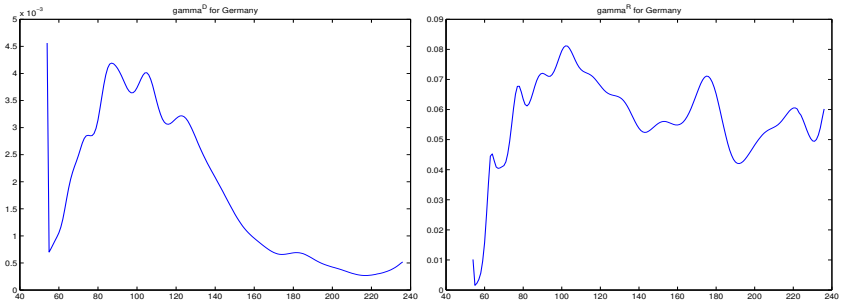

Fig. 27 Time series for estimates of γ^D and γ^R for Germany up to day 236

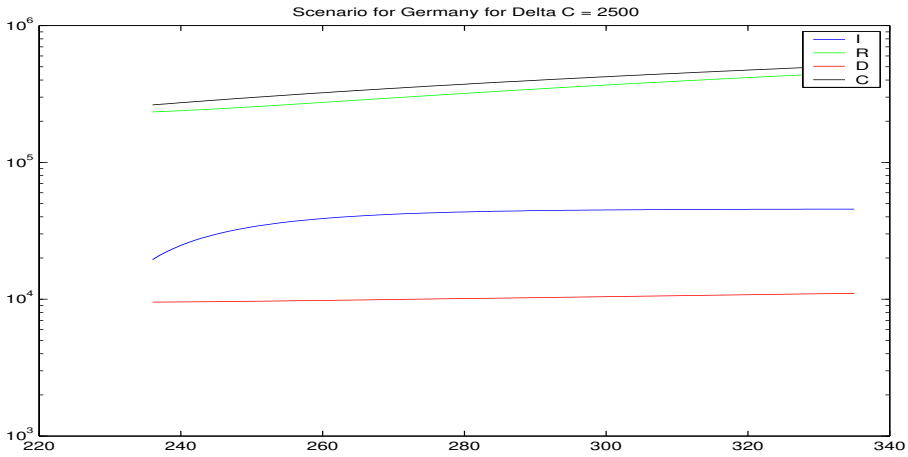

Fig. 28 Prediction for $\Delta C = 2500$ for Germany based on data up to day 236

7.5 Old Predictions of July 2020

Since early June, new superspreading events took place in Germany, and the Infectious were increasing again, but decreasing after fast intervention by the administration. Various countries seem to have a second outbreak. This calls for updates of the figures in the paper. But also, the data quality gets worse over time, and predictions get more difficult. For example, the UK figures of Recovered always were useless, but on July 19th, even the registration of deaths will be suspended, making any predictions impossible.[3]Data of other states are suspected to be not only questionable, but even manipulated.[4][5]

[3] July 19th: https://www.independent.co.uk/news/uk/politics/coronavirus-uk-death-toll-nhs-phe-covid-19-government-england-scotland-a9626336.html

[4] June 9th: https://www.telegraph.co.uk/news/2020/06/09/coronavirus-world-round-up-jair-bolsonaro-accused-manipulating/

[5] May 19th: https://kutv.com/news/coronavirus/states-accused-of-manipulating-covid-19-statistics-to-make-situation-look-better

Country	Death rate	Detection rate	Death rate	Detection rate
Germany	0.039	0.128	0.017	0.300
Brazil	0.040	0.125	0.039	0.170
Italy	0.117	0.043	0.074	0.068
Spain	0.147	0.034	0.017	0.300
Sweden	0.033	0.151	0.023	0.215
Austria	0.041	0.123	0.011	0.449
France	0.083	0.060	0.054	0.092
UK	0.114	0.044	0.133	0.038
US	0.036	0.141	0.014	0.346

Table 4 Case fatality and detection rates, estimated on day 159 (June 29th) and day 179 (July 19th) using the 14-day rule and a backlog of 28 days.

7.5.1 Predictions

The full model predictions replacing Figure 8 are in Figure 29, now based on data up to day 179 (July 19th). Germany has overcome the small intermediate increase of the registered Infectious I (black □) and the hidden Infectious M (cyan □) around day 150-160 due to outbreaks e.g. in the Tönnies factory on day 146. But lowering the restrictions has stopped the rapid decline around day 100 and replaced it by a very gradual decline. The same is visible for France, and in the plot for US one can see that the increase of Infectious got larger. Anyway, Germany still has good chances to stay below a total of 10,000 deaths, while the long-term prediction of the full model for the US has an R_0 of 1.93 and targets a final death toll of 2 million with a peak of the Infectious around day 380 (Feb. 5th, 2021) if no actions are taken. Brazil also showed a strong increase, but was taken out due to unreliable data. Spain was included to show how fast the model reacts to new small outbreaks. But the results for deaths are very questionable shortly after a new increase of Infectious, because the instantaneous case fatality rate is near to zero until the deaths following the new infections show up. This applies to Spain and Germany. For reasons to be explained below, the results for France are somewhat questionable.

The update of Table 1 is Table 4, based on data up to day 159 (June 29th) and day 179 (July 19th), using the backlog of 28 days. Most death rates are smaller now, and detection rates are mostly higher. The Recovered of Sweden and the UK had to be estimated using the 14-day rule by [25].

For readers interested in what happened to Figure 11 describing the COVID-19 situation in Göttingen, here is Figure 30. The two superspreading events in rundown apartment houses are clearly visible, and how the local authorities regained control.

8 Mathematical Additions

These were parts of earlier additions to the basic paper [24]. We state them here close to their original form. Some of them were implemented for the new computational results in section 7.

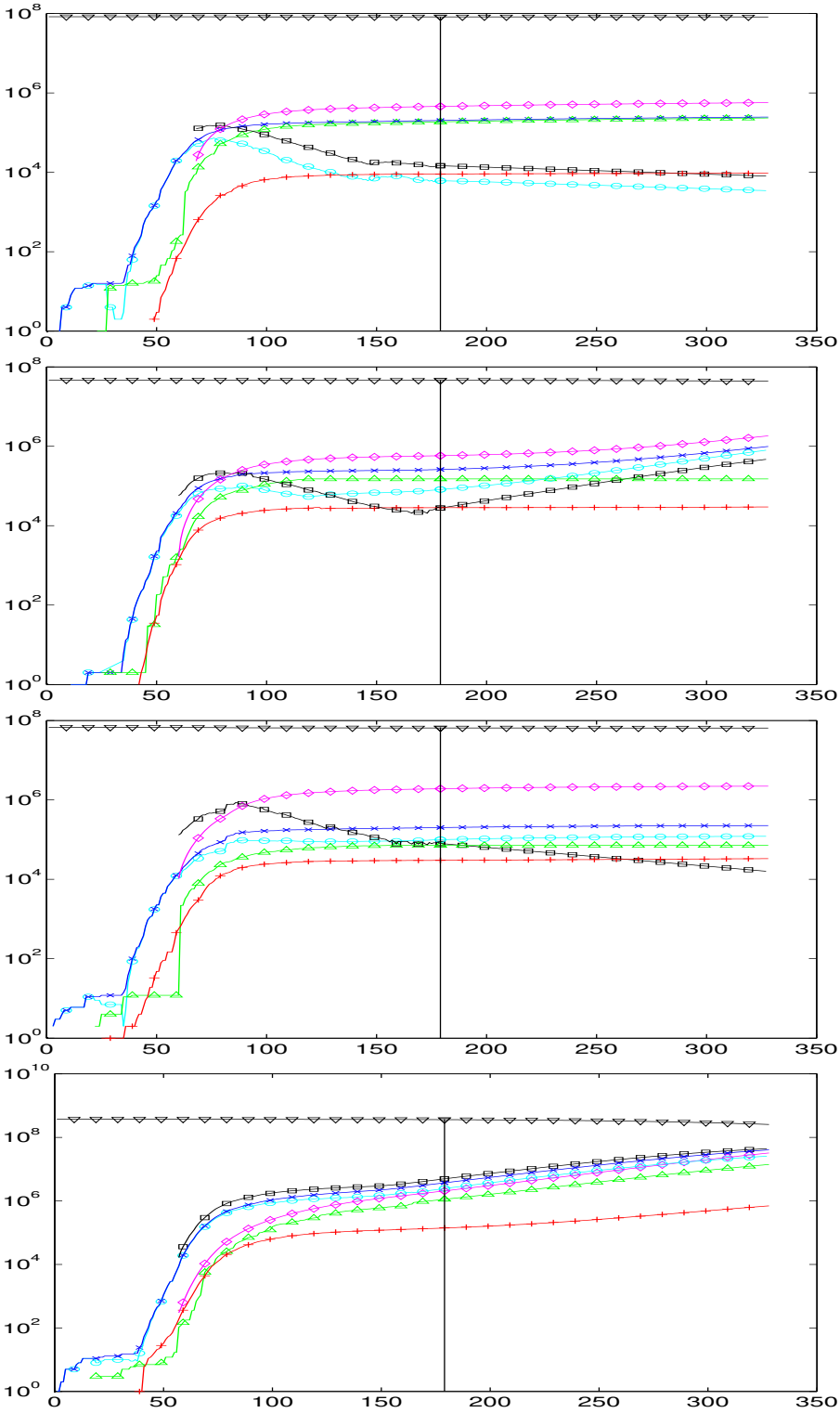

Fig. 29 Predictions for countries Germany, Spain, France, and US on day 179 marked by the vertical line. The S, M, H values to the left are obtained by calibration, the C, R, D, I values there are the original Johns Hopkins data. Not all data points have marks. See Figure 8 for explanation of colours and markers.

Fig. 30 Infectious in Göttingen city and county, as of end of July 2020 in the local newspaper "Göttinger Tageblatt". Still no exponential outbreak in the large, but two local superspreading events.

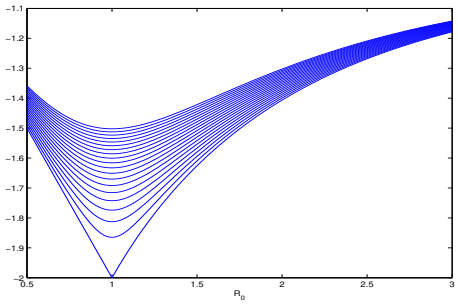

Fig. 31 The asymptotic decay rate (40) as a function of R_0 for $s(0) = 0.99^k, k = 0,1,2,\ldots$, from below.

8.1 Additional Explanations

Here, some additional remarks are collected to improve readability of certain parts of the paper.

8.1.1 Asymptotic Exponential Decay

This concerns the decay rate

$$R_0 s_\infty - 1 = W(-s(0)R_0 \exp(-R_0)) - 1 \tag{40}$$

in (16), and the MAPLE-based statement there that this is smallest for $R_0 = 1$ no matter how large $s(0)$ is. If one takes the derivative wrt. R_0, one sees (and MAPLE confirms) that $R_0 = 1$ is the only zero. See Figure 31 for several cases, the lowest one for the impossible $s(0) = 1$.

8.1.2 Peak Location

This concerns section 2.14 and the remarks made there concerning MAPLE programs. We start right after the upper bound (20) and explain the MAPLE argument for the lower bound leading to (21).

The goal is to find combinations of $s(0)$ and R_0 satisfying the peak condition (7) such that the integral in (19) gets arbitrarily large. This will deal with values extremely close to one in (7), and a peak that is extremely flat and far away. Therefore it is reasonable to replace i in the integral (19) by its maximum (18) at the peak. Then we get the lower bound

$$\begin{aligned}\tau_I &\geq \frac{\log(s(0)R_0)}{R_0}\frac{1}{1-\frac{1}{R_0}-\frac{1}{R_0}\log(s(0)R_0)}\\ &= \frac{\log(s(0)R_0)}{R_0-1-\log(s(0)R_0)}.\end{aligned}$$

Instead of explaining how the paper proceeds at this point, we follow a much nicer approach based on a result of Wolfgang Hensgen, stating that in a peak situation (7) the specific initial condition

$$s(0) = \exp\left(1-\frac{1}{R_0}-\log(R_0)\right) < 1$$

satisfies the requirements and yields

$$\tau_I \geq \frac{1}{R_0-1}. \tag{41}$$

One can improve this slightly when setting $z := \log(R_0 s(0))$ to get

$$\tau_I \geq \frac{z}{R_0-1-z}$$

which coincides with Hensgen's result for $z = 1-1/R_0$. This lower bound is monotonic in z, and since z is confined to $0 < z < \log(R_0)$, the largest lower bound arises for the limit $z = \log(R_0)$ or $s(0) = 1$ with

$$\tau_I \geq \frac{\log(R_0)}{R_0-1-\log(R_0)} \geq \frac{2}{R_0-1},$$

the latter inequality obtained by MAPLE. Anyway, for each fixed $R_0 > 1$ the lower bound (41) by Hensgen holds for the range

$$\frac{1}{R_0} < \exp\left(1-\frac{1}{R_0}-\log(R_0)\right) \leq s(0) < 1$$

of starting values $s(0)$.

With the above substitution $z := \log(s(0)R_0)$ we now compare upper and lower bounds to find

$$\begin{aligned}
\frac{r(\tau_I)}{i(\tau_I)} \le \quad & \tau_I \quad \le \frac{r(\tau_I)}{i(0)} \\
\frac{z}{R_0 - 1 - z} \le \quad & \tau_I \quad \le \frac{z}{R_0 - e^z} \\
1 \le \tau_I \frac{R_0 - 1 - z}{z} & \le \frac{R_0 - 1 - z}{R_0 - e^z} =: g(z, R_0)
\end{aligned}$$

showing that bounds get sharp when $s(0)R_0$ comes close to one, even when R_0 is large. For large fixed R_0 and initial conditions $s(0)$ close to the minimal $1/R_0$ needed for a peak, the peak moves to zero roughly like

$$\tau_I \approx \frac{\log(s(0)R_0)}{R_0 - 1}.$$

This handles the cases where the two bounds are close. But they are far apart in the most interesting case: a fixed $s(0)$ very close to 1 and a varying $R_0 > 1/s(0)$. Figure 32 shows the peak location τ_I as a function of $R_0 - 1$ for different $s(0) = 1 - 2^{-k}, k = 1, \ldots, 40$ starting from the lowest curve. The derivative singularity at $R_0 = 1$ was resolved by a logarithmic scale of $R_0 - 1$, and the integral was handled by the recursive adaptive Lobatto quadrature routine `quadl` of MATLAB.

We know from section 2.14 that the curves go to zero at both ends. It is clear that the peak moves out when $s(0)$ moves closer to one, but it is strange that R_0 near to the Euler number e seems to have the maximum of the peak position for $s(0)$ close to 1. For $s(0) = 1 - 2^{-k}$ one roughly has something like

$$\tau_I \approx \alpha k \log(R_0 - 1) + c$$

for $R_0 \le 2$ and large k, because the left parts of the functions in Figure 32 have the same derivative and step up linearly with increasing k. This may be a rule-of-thumb for small $i(0) = 1 - s(0)$ and $1 \le R_0 \le 2$, but the full determination of the peak location is still open.

This and the next section were significantly improved by e-mail contact to Wolfgang Hensgen (Munich).

8.1.3 Global Existence of Solutions

As correctly criticised by Wolfgang Hensgen by e-mail, the paper does not address the question of global existence and positiveness of solutions of the SIR model. This is serious, because the right-hand side of (2) is quadratic, and the example $y'(x) = -y^2(x)$ with singular solutions $y(x) = 1/(x - c)$ requires to exclude singularities, for instance.

We start with a geometric solution that sheds some additional light on section 2.11. A shorter solution by Wolfgang Hensgen follows below. The solution vectors $(r(\tau), s(\tau))$ of the SIR system (2) should be in the triangle

$$M := \{(r,s) \in [0,1]^2,\ s + r \le 1\},$$

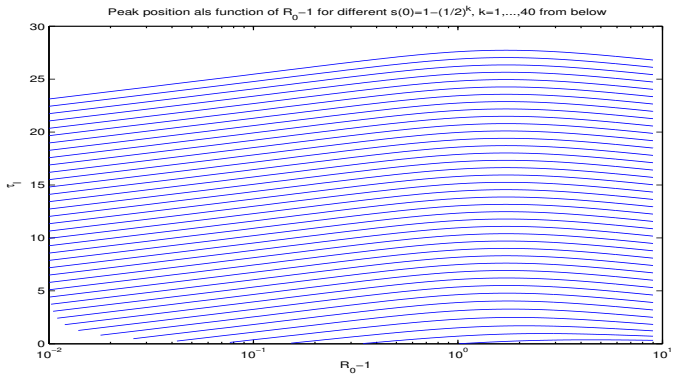

Fig. 32 Peak location τ_I as a function of $R_0 - 1$ for different $s(0) = 1 - 2^{-k}$, $k = 1, \ldots, 40$ starting from the lowest curve.

and we consider SIR solutions starting in points $(r(0), s(0)) \in M$ with $0 < s(0) < 1$ to avoid trivial cases.

But we take a detour via the unrestrictedly solvable system (4) under addition of

$$i(r) := 1 - r - s(r) = 1 - r - s(0)\exp(-R_0(r - r(0)))$$

and follow the solutions as curves $r \to (r, s(r))$ in M, like in Figure 33. Without loss of generality we can work with $r(0) = 0$, extending curves to the left, if necessary. The triangle below the red line is M, and all curves hit the red line $s = 1 - r$ as soon as $i(r) = 1 - r - s(r)$ above is zero, and this occurs at $r_\infty < 1$ defined by (15) with Lambert's W function.

Each piece of the possible $r \to (r, s(r))$ curves in the interior of M coincides as a point set with a piece of a SIR solution curve $\tau \to (r(\tau), s(\tau))$, and vice versa. The two curve parametrizations are connected by

$$\frac{d\tau}{dr} = \frac{1}{i(r)} = \frac{1}{1 - r - s(0)\exp(-R_0 r)}$$

This proves that all SIR solutions starting in M with $s(0) < 1$ will stay in M and not get singular.

But it is not yet clear why τ is not bounded. The right-hand side of the SIR system (2), when restricted to M, is uniformly Lipschitz continuous for each fixed R_0. By standard ODE solvability theory, this implies that the domain of each solution can always be extended by a fixed $\Delta\tau$ depending only on R_0. Therefore the SIR system is globally solvable and the limit for $\tau \to \infty$ of a curve $\tau \to (r(\tau), s(\tau))$ is $(r_\infty, 1 - r_\infty)$ on the boundary of M. We have used this in section 2.11 of the paper.

As an aside, all SIR solutions can be obtained as point sets by calculation "back from infinity" by picking some $R_0 > 0$ and some $r_\infty \in (0, 1)$ with the constraint

$$(1 - r_\infty)\exp(R_0 r_\infty) < 1.$$

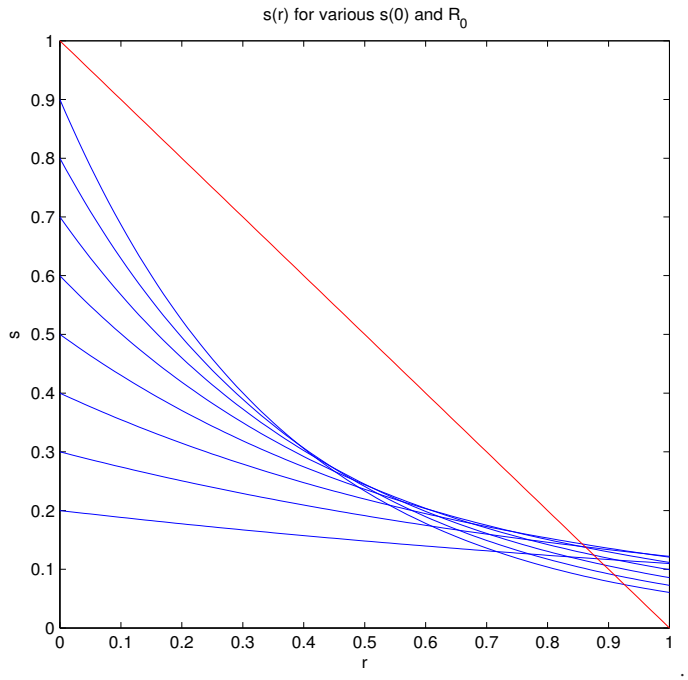

Fig. 33 Curves $s(r)$ for various $R_0 > 0$ and $s(0) \in (0,1)$

Setting $s_\infty := 1 - r_\infty \in (0,1)$ and solving (4) backwards from r_∞ to $r = 0$ will end up with the above inequality as $s(0) < 1$. This traces the point sets of the curves in Figure 33 from right to left, starting at the red line. However, the τ parametrization is not recoverable this way.

We finish by presenting a shorter proof due to Wolfgang Hensgen. As long as s and i are positive, solutions of the system (2) will satisfy the nonlinear system of integral equations

$$\begin{aligned} s(\tau) &= s(0)\exp\left(-R_0\int_0^\tau i(x)dx\right) \\ i(\tau) &= i(0)\exp\left(\int_0^\tau (R_0 s(x) - 1)dx\right), \end{aligned} \tag{42}$$

and we follow solutions of (2) and (42) started on values $s(0), i(0) \in (0,1)$ with $r(0) = 1 - s(0) - i(0) \in [0,1)$ on the domain

$$M := \{(s,i) \in [0,1]^2,\ s + i \le 1\}$$

for τ in $[0,\tau_0)$, where we admit $\tau_0 = \infty$.

By (42), solutions cannot reach $s = 0$ or $i = 0$ for finite τ, and due to $r' = i$, solutions cannot reach $s + i = 1$, i.e. $r = 0$. But since the ODE system (2) has a uniformly Lipschitz continuous right-hand side on M, solutions can be uniquely continued for

$\tau \in [0,\infty)$, proving global unique existence of solutions, staying in M. As we know from Section 2.11, the boundary of M is reached for $i(\tau) \to 0$ with positive limits from (13) for the monotonic functions r and s at infinity.

8.2 Constant New Confirmations

A typical situation in Germany in mid-July 2020 was that there was a roughly constant number ΔC of new Confirmed, varying between 200 and 500, and slowly increasing towards 2000 in mid-September. This can be due to more travelling or relaxation of contact constraints, for instance, but here we do not ask for reasons. Instead, we check how the Johns Hopkins data model behaves under such an assumption. Equations (23) turn into

$$\begin{aligned} C_{n+1} - C_n &= \Delta C, \\ I_{n+1} - I_n &= \Delta C - \gamma_n I_n, \\ (R+D)_{n+1} - (R+D)_n &= \gamma_n I_n \end{aligned}$$

maintaining the balance $C = R + D + I$ between the cumulative Confirmed C, the cumulative Recovered R, the cumulative Deaths D, and the non-cumulative Infectious I. If the γ_n are considered to be constant, the I values go exponentially to a level $I = \frac{\Delta C}{\gamma}$, from below or above, depending on the starting value for I. The qualitative behaviour of I now is logistic. There is no peak as in section 2.8 and no long-term exponential decrease to zero as in section 2.12. But a constant number of Infectious means a constant increase of deaths, calling for political changes that stop this scenario. Anyway, it cannot be valid over long time intervals because there is an upper bound on the Confirmed.

To get a grip on deaths and recoveries, we should split the last equation into

$$\begin{aligned} \gamma_n^R + \gamma_n^C &= \gamma_n, \\ R_{n+1} - R_n &= \gamma_n^R I_n, \\ D_{n+1} - D_n &= \gamma_n^D I_n. \end{aligned}$$

The γ_n^R and γ_n^D are now time series of the instantaneous case rates γ_{iCR} and γ_{iCD} of (26), and it is easy to get estimates of them by just solving the equations. Figure 34 shows these values as time series, based on the Johns Hopkins data after applying the technique of section 8.4.3. For Germany on day 188 (July 28th) using a mean over the last 14 days, the values are

$$\gamma^D \approx 0.0005,\ \gamma^R \approx 0.064.$$

With a number of about $I = 7600$ Infectious on day 188, the critical value of ΔC determining increase or decline of I is $\Delta C = I\gamma \approx 500$. Then a constant I of about 7600 implies a daily death count of four due to COVID-19. This agrees quite well with what actually happened in mid-July, but this must be expected because we used the real data. If rates γ^R and γ^D stay as they are, doubling ΔC implies doubling the asymptotic I level and the death rate in the long run. The logistic increase to the new I level is rather fast, see Figure 35 for Germany with an assumed $\Delta C = 1000$ from

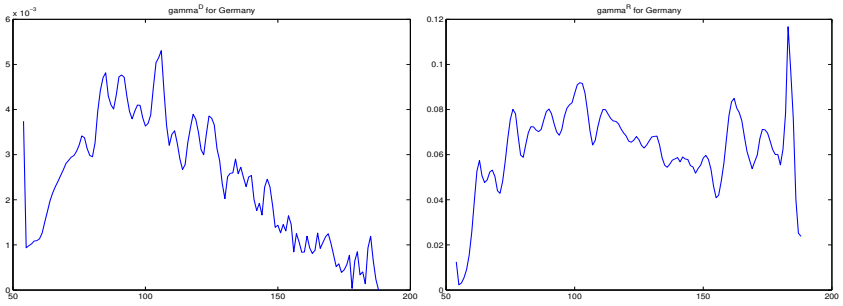

Fig. 34 Time series for estimates of γ^D and γ^R for Germany up to day 188

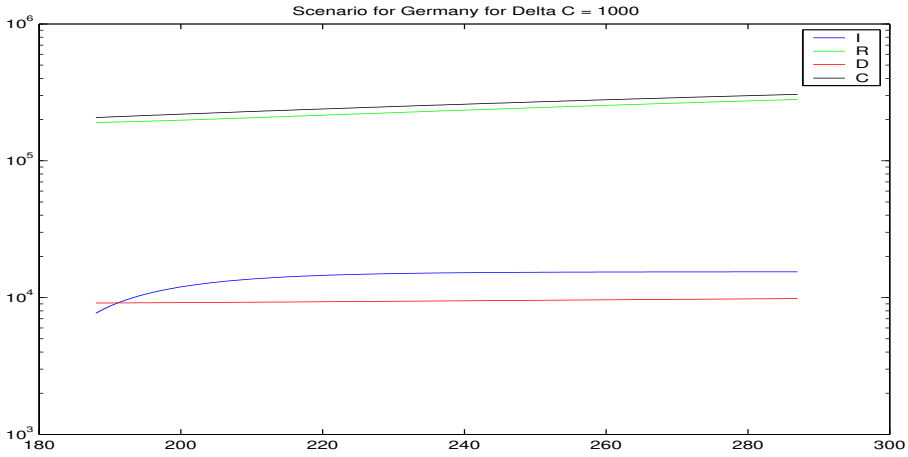

Fig. 35 Prediction for $\Delta C = 1000$ for Germany based on data up to day 188

day 188 on. The linear increase of C is still far from reaching saturation, and the total death toll still stays below 10,000 in the period shown.

Italy has about 250 new daily infections around day 188, with values of $\gamma^D \approx 0.00086$ and $\gamma^R \approx 0.124$ with Infectious around 12,350. The threshold for ΔC is around 125, and a value of 400 new daily infections would imply about 11 daily COVID-19 deaths in the long run.

Summarizing, there are situations where countries with a small and constant daily increase ΔC of the Confirmed may get along for quite some time, but at the price of a constant and hopefully low COVID-19 death rate. Societies and their politicians must decide whether this rate is tolerable.

But various other countries like France, Spain, UK, or US cannot be treated that way, as of mid-July 2020, because either the current ΔC values are not roughly constant or the γ values are completely unreliable because the R or D values are. Finally, the γ^D values used in the above examples were much smaller than what was observed in the early phase of the COVID-19 outbreak, see Figure 34. If health systems get under serious stress, and if seniors are not properly protected, the above scenario will become much worse.

8.3 Shortcomings of the Models

Here, we list several effects that turned up after publication of the basic paper [24] and revealed certain problems when running both the Johns Hopkins Data Model of 3.3 and the full model of section 5 behind a peak, at low numbers. Section 10 provides details about changes to programs that take care of these shortcomings.

But before we turn to the models, we list shortcomings of the data needed by the models. There should be external time- and country-dependent additional data on

- infection fatalities,
- case fatalities,
- average case histories,
- detection rates and testing,
- subgroups, e.g. age groups.

These are partially available, but only for specific times and countries. Data-driven models like these here suffer from unsafe or missing data, and will greatly improve when there is more useful information.

8.3.1 Starting Values and Delays

The models in the paper rely strongly on the numbers for the deaths, when it comes to estimating the hidden variables or the case fatalities. This has serious consequences for re-infections after peaks. As long as only the Infectious re-increase without any influence on deaths and recoveries, the new situation does not have a strong influence on the hidden system. In particular, re-infections will quickly increase the R_0 estimate by the Johns Hopkins Data Model, but will only reluctantly increase the R_0 estimate by the full model, because the hidden part does not change as quickly.

This is one of the places where a major shortcoming of the models shows up: the inadequate treatment of delays. Another case is that the prediction of a second peak should work like a restart of the full model under new initial conditions. But like startup data are discarded until at least 10 deaths and 100 Confirmed are present, the restart requires again at least 10 new deaths and 100 new Confirmed. It cannot be expected that the models produce useful results unless this condition is satisfied.

8.3.2 Estimating R_0 from Johns Hopkins Data

Some formulas rely strongly on monotonicity of the data, e.g. (24), (29), and (32). And, when cumulative time series get near-stationary after a peak, the R_0 estimation of the Johns Hopkins Data Model by (24) comes close to 0/0 and thus becomes extremely unstable. In particular, the estimation by (24) will get very large when the Confirmed already re-increase while deaths and Recovered do not yet follow up. See data of France in Figure 19 for a typical situation.

The naive way to repair this is to form the time series for r_n using (24) after applying the technique of section 8.4.3 for guaranteeing monotonicity and a limited amount of smoothness. This was still too unstable, and a way around the 0/0 effect for small values of new deaths, recoveries, and confirmations was needed. The current workaround is described in section 7.4.2 and uses the derivative of $C(R+D)$.

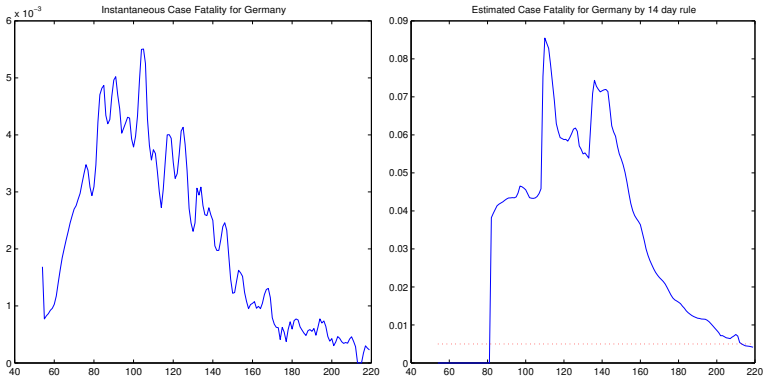

Fig. 36 Instantaneous and estimated case fatality rates as time series, for Germany up to day 219 (August 28th). The dotted line marks the infection fatality rate of 0.005 used in earlier versions.

8.3.3 Replacing Constants by Time Series

Experience with the models shows that nothing should be considered as being constant. Therefore one should use time series instead of constants, wherever possible. However, for predictions of the future, one has to go back to constants or implement laws that provide time series for "constants" in terms of variables in the simulation. The latter has not yet been implemented, but time series were introduced in the calibration phase of section 4.5. For the paper, this phase was rather short and coincided largely with the infection peak, allowing the use of constants, but now the time variation of "constants" has to be accounted for.

8.3.4 Time-dependence of Case Fatalities

The strongest deviation from constants concerns case fatalities. The version of July 29th already used "startup" values for (33) in model calibration, but "final" values in the full model for prediction. The software for the DMV paper always used the "final" values.

But a closer inspection of the data reveals that case fatalities should be time series, not just one or two constants. Nearly all countries were quite successful in reducing their case fatalities behind the peak, with a roughly linear reduction following by a level-out that comes close to or even significantly below the infection fatality of 0.005 assumed so far. This calls for a lower infection fatality rate as well, see section 8.3.5.

The background of the decline in fatality rates may be that detection rates got better by extended testing, leading to more confirmed cases with much less serious symptoms. But there may also be more confirmed cases with better health condition, once the peak is over and the endangered persons are better protected. Another reason may be that the health systems improved their treatment strategies. The literature [21], [27],[28] does not contain helpful new information on fatality and detection rates and their variation over time until recently.

To show these variations in fatality rates, Figure 36 shows the observable instantaneous case fatality rate γ_{iCD} from (29) for Germany together with the estimated case fatality rate γ_{CF} using the 14-day rule and based on data 28 days backwards, calculated as in section 4.3.2. The reduction in case fatality rates over time is clearly visible, the estimated value being roughly a multiple of 20 of the instantaneous value. The rough agreement in the two plots of Figure 36 confirms that the estimation technique of section 4.3.2 for case fatalities is consistent with the available data on instantaneous case fatalities.

Anyway, the consequence for data-driven modelling is that case fatalities should enter as time series, as far as possible. The most important places are (31) and (33), making the factors in both equations of (34) time-dependent and leading to a serious change in the calibration process of section 4.5. The case fatality rate γ_{CF} is now calculated time-locally via section 4.3.2, not globally. The k-day rule is applied for $k = 14$ using data for $2k$ days backwards.

8.3.5 Time-dependence of Infection Fatalities

But the case fatality now drops below the global constant 0.005 used so far for the Infection Fatality Rate γ_{IF}. One way out is to use a much smaller constant γ_{IF}, but this is unrealistic when modelling the early exponential outbreaks in spring 2020. In the models, it makes detection rates (30) extremely small and leads to extraordinarily high percentages of hidden Infectious. Furthermore, detection rates should be time series as well, because testing strategies change. Thus the infection fatality should also be a time series, but which?

There are a few requirements. During peak time, the infection fatality should be allowed to be large, e.g.$\gamma_{IF} = 0.005$ as in the simulations for the paper, based on experimental results for spring 2020 cited in section 4.3.1 of the basic paper. But it should be allowed to drop when the case fatality drops, in a similar fashion. However, the detection rate should be bounded above by some value $\delta \leq 1$ and increase when the case fatality decreases, to mimic improved testing when fighting the pandemic successfully.

There are many ways to satisfy these restrictions, and the results do not differ much in the final model results, qualitatively. After quite some experimentation, the following model change was implemented:

$$\gamma_{n,IF} = \gamma_{IF}(1 - \exp(-\delta\gamma_{n,CF}/\gamma_{IF})) \tag{43}$$

defines a time series for the infection fatality rate as a function of the time series for the case fatality rate. Figure 37 shows the behaviour of this function for $\gamma_{IF} = 0.005$ and $\delta = 0.5$, together with the implied detection rate. The infection fatality rate starts from 0.005 and gets smaller with a smaller case fatality rate, while the detection rate improves gradually, up to $1/2$ in the limit situation without any fatalities. The data of Table 1 are marked, and they are close to the curve because the table was calculated for $\gamma_{IF} = 0.005$, and the case fatalities were large at that time. Predictions in September 2020 still use $\gamma_{IF} = 0.005$ to start with, but set a maximal detection rate of 20% for Table 3.

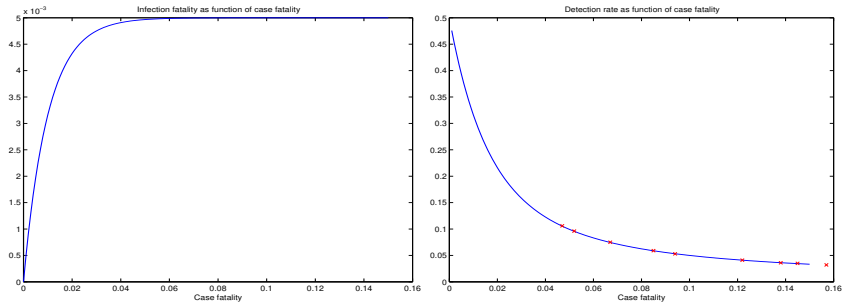

Fig. 37 The infection fatality (43) and the detection rate $\gamma_{n,IF}/\gamma_{n,CF}$ as functions of the case fatality $\gamma_{n,CF}$.

8.4 Imputation and Nowcasting

This concerns improvements of the data quality by reducing errors induced by delays in data transmission. We first sketch how the Robert Koch Institute deals with this problem, and then we provide a somewhat similar technique to improve the Johns Hopkins data.

8.4.1 Methods of the Robert Koch Institute

In [22] of May 15th, the Robert Koch Institute published its current way of estimating R_0. This is based on the Imputation and Nowcasting data corrections described by An der Heiden and Hamouda in [12]. It roughly implements the logic of (11) because it takes delayed ratios of Infectious and applies additional means. This is not estimating the R_0 of the mathematics of the SIR model. It estimates R_t in the formulation of Section 2.5 of the paper, namely as the multiplication factor for new Infections, restricted to registered cases. This factor is quite sufficient in practice, since for the German situation the impact on the health system is very strongly connected to the new daily registered infections. See section 8.2 for a variation of the Johns Hopkins Data Model dealing with this situation and making predictions.

Imputation by the RKI concerns the estimation of a fictitious symptom onset from the known registration time, i.e. an estimation of the delay between symptom onset and registration. The RKI found a delay between 5.3 and 9 days, time-varying, and applied this to the 40% of all cases where the symptom onset was not known. For about 60% of the data, the RKI knows the delay. To do this properly for the Johns Hopkins data, one needs information about the correspondent delays in Johns Hopkins data. These are missing, to my knowledge, and will anyway be strongly country-dependent.

Nowcasting corrects a time series that counts "cases", when the assignment of the cases to the times is doubtful, e.g. due to stochastic delays in transmission. If a doubtful case count x_j is observed at time j, it may contain true cases X_i at time i for $i \le j$ by delayed transmission. Conversely, true values X_n contain parts of the x_k for $k \ge n$. If q_j is the portion of cases that should have been registered j steps earlier, one

has

$$X_n \approx \sum_{j=0}^{D-1} q_j x_{n+j}. \tag{44}$$

The RKI seems to use imputation to assign the "right" n to uncertain x cases to account for the correct infection day, and then use the obtained x_n for nowcasting. To do this properly one needs experimental statistics about observed delays, or Bayesian priors to assume meaningful distributions of such delays. The RKI has such information, but what about Johns Hopkins data?

From a deterministic viewpoint, (44) will have a smoothing effect, and this may be a way to use it when no additional information is given. We shall apply (44) in section 8.4.3 to mimic imputation and nowcasting without having additional data on delays.

8.4.2 Repairs in Johns Hopkins Data

The Johns Hopkins data contain cases where cumulativity is violated, and where strong ≈ 7 day oscillations occur that may, like in the RKI case, be due to delays in transmission, e.g. by weekends. Furthermore, there are violations of the k-day rule, and predictions will improve if certain conservative smoothing techniques are applied. "Conservative" means here and in the basic paper that counts of cases are kept in the mean. Missing or doubtful data for Recovered can be roughly estimated using [25] and a k-day rule, if the Confirmed and Deaths are reliable. A general goal is to find algorithms that detect errors and plausibility flaws in the Johns Hopkins data, but this is ongoing work, and we start with a simple case.

8.4.3 Monotonicity and Smoothness

Assume that a time series X_n is required that is at least weakly monotonic, based on observations x_n that may be faulty. The goal is to find an algorithm that works like imputation and nowcasting, combined with a smoothing technique, and guarantees monotonicity. This will be applicable to C, R, and D of the Johns Hopkins data.

The idea pursued here is to perform a fit like (44) under monotonicity constraints and non-smoothness penalties. For the latter, we may measure smoothness by a vector-valued function f whose components $f_j(X)$ give a penalty for non-smoothness of the X values around step j. One way is to use a time window around time j, take the X_k in the window, fit them to a fixed local model and return the fitting error. Then one might minimize a weighted sum of $f_j(X)^2$ and squared errors in (44) under all X with monotonicity constraints and under all nonnegative q_j. If necessary, one can restrict the sum of the q_j to be one, enforcing that no cases are lost in the mean.

The implementation works via MATLAB's `lsqnonlin` acting on the logarithms of C, R, or D, the non-smoothness penalties being the fitting error to a quadratic polynomial on a window of up to five points. The overall optimization problem thus is quadratic with linear constraints, usually leading to a unique solution. If M data are

j	C	R	D
0	0.2085	0.0949	0.1261
1	0.1985	0.1369	0.1393
2	0.1965	0.1761	0.2284
3	0.1284	0.1907	0.1577
4	0.1090	0.1738	0.1922
5	0.1167	0.1326	0.0878
6	0.0428	0.0924	0.0685
sum	1.0004	0.9975	0.9998

Table 5 Quantities q_j for monotonizing, nowcasting and smoothing the Johns Hopkins data of France up to day 170. There was no constraint on the sum of the q_j.

treated, this minimizes over the $D+M$ nonnegative quantities (q,X) under monotonicity constraints on X and $2M-D$ penalties consisting of $M-D$ errors in (44) and M non-smoothness penalties. Note that the non-smoothness penalties are independent of the observed data. A more sophisticated technique of non-smoothness penalties will not aim at local quadratics, but could fit data on 5 points by two line segments with a breakpoint at one of the points. This is currently not implemented, but would not iron out any breakpoints. Note that equation (5) in Höhle/an der Heiden [13] uses a local quadratic spline for smoothing within a Bayesian nowcasting technique.

For Johns Hopkins data of France up to day 180 (July 20th), the C, D, and R values have 20, 7, and 12 places of non-monotonicity, respectively. Figure 38 shows the result of the algorithm for $D = 7$. Here and in many other cases, the new data (dotted) are to the left of the original data. This is normal, because they contain values that are possibly falsely registered on later days. Table 5 shows the resulting q_j for the data of France. They indicate that Recovered are possibly registered later and more irregularly than Confirmed and Dead. This calls for further checks, if time permits.

For countries with useless data for the Recovered R, like the United Kingdom, the Confirmed C and Deaths D data are processed first, and then the R data are estimated from the processed C and R data under the 14-day rule as proposed in [25].

For calculations from September 2020 on, an additional non-smoothness penalty of the same type was applied to differences, and a normalization by dividing through the local maximal value was applied to the smoothness penalties. Finally, the last $2D$ values got an additional final $(1/4,1/2,1/4)$ smoother. The reason was that (44) cannot treat the last D values of n, and those $\hat{X}_n$ were directly fitted to the x_n, leading to a derivative jump exactly D days before the end of the time series. This effect produced strange kinks in plots like Figure 23, and required some ironing. In addition, to play safe, the sum of the q_j was forced to be close to one. See Figure 19 for the results of the improved algorithm.

8.5 Testing the k-day Rule

A standard assumption is that the newly Confirmed $C_n - C_{n-1}$ at day n will end up dead or alive until day $n+k$ for k large enough. This was called the *k-day-rule* in

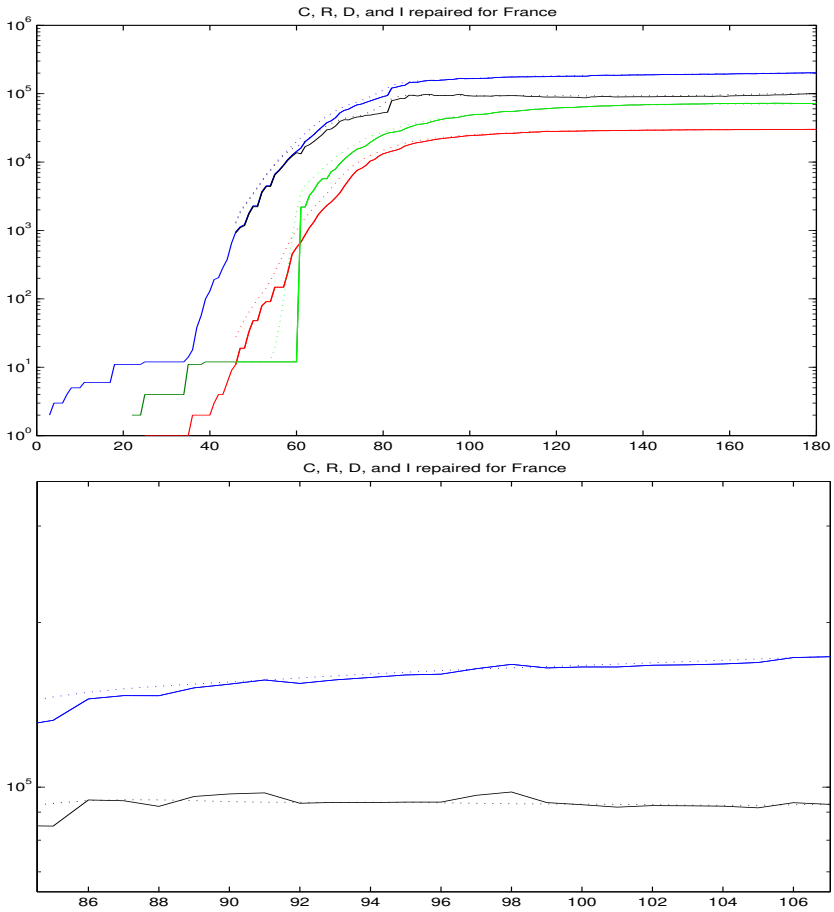

Fig. 38 The Johns Hopkins data for France, with repair, up to day 170 (July 20th). Dotted: Repaired data. Full data and a closeup. Note that the black I curves are not monotonic.

the paper, and this section intends to give some supporting information. There is no direct influence of this section on the models.

The k-day rule implies

$$C_n - C_{n-1} \leq R_{n+k} - R_{n-1} + D_{n+k} - D_{n-1}, \tag{45}$$

because all newly Confirmed must be dead or recovered k kdays later. Conversely, the new Removed $D_n - D_{n-1} + R_n - R_{n-1}$ at day n must have been confirmed between day n and day $n-k$. This means

$$D_n - D_{n-1} + R_n - R_{n-1} \leq C_n - C_{n-k-1} \tag{46}$$

and both inequalities admit that transition from Confirmed to Removed can be on the same day. This is different from the treatment of the k-day rule in the paper [24] and

in [25], where death and recovery can occur at least one day later than confirmation. Altogether, this differs by just a shift by one day. The extreme case $k = 0$ now implies that

$$C_n - C_{n-1} = R_n - R_{n-1} + D_n - D_{n-1},$$

i.e. all new Registered at day n are either dead or recovered. Since the k-dependent right-hand sides increase with increasing k, there should be a minimal k_{min} for which the above relations hold for all reasonable n.

Table 6 shows the minimal k for which either (45) or (46) are satisfied for the full range of data with $D \geq 10$ and $C \geq 100$ up to day 179 (July 19th). The data were preprocessed with the technique of section 8.4.3 in the July version. Furthermore, for UK, Denmark, and Sweden, the Recovered had to be estimated by the technique of [25] to get reasonable values at all. If the calculation is repeated for the data between days 142 and 170, all values (except for Spain) are considerably smaller. The unusually high value for Spain is due to failing (45) by at most 300 cases out of about 250,000 Confirmed. Altogether, even for the early phase of the outbreak, a 14-day rule is roughly satisfied, supporting the strategy used so far for estimation of Recovered and of the case fatality rate.

Country	Passes (45) for k	Passes (46) for k
Austria	12	4
Brazil	18	1
Denmark	15	3
France	13	1
Germany	13	1
Italy	6	5
Russia	12	1
Spain	29	4
Sweden	15	1
Switzerland	14	4
UK	13	1
US	11	1

Table 6 The minimal k for which the Johns Hopkins data up to day 179 satisfy either (45) or (46).

9 Non-Mathematical Additions to the Text

9.1 Media Coverage in German TV

As of early July, the standard broadcasts does not mention terms like Reproduction Number or Doubling Time anymore. Mostly, they give the 7-day mean of *new* infections and the increase or decrease of it. This is information about the second derivative of the Confirmed, and therefore useful, though it would be better to report the actual Infectious and their derivatives. The strong variations in the data provided by the Robert Koch Institute are smoothed away by the 7-day mean that is often mentioned,

but not always. Daily values are only rarely given, avoiding any false positive interpretation of lower values on Mondays due to the transmission delays incurred by the weekend. The data changes by the RKI, as roughly described in section 8.4, are not mentioned, and the number of currently infectious persons is only rarely provided.

9.2 RKI Papers

From April 9th on, the Robert Koch Institute (An der Heiden & Hamouda [12]) published its own way of preprocessing its data by *Imputation* and *Nowcasting*. I missed this publication when working through the RKI website in March, and got it in June via Dehning et al. [5]. This led to section 8.4 for implementation of a similar strategy in 8.4.3.

9.3 Excess Mortality

In general, the data of August/September have the common feature that death rates are much smaller than in the days of the exponential outbreak. The estimated case fatalities come down to values around 0.002, much lower than the infection fatality 0.005 used in earlier predictions and the observed 0.0036 by Streeck et al. [29]. There may be many reasons for this, but we have to take the data as they are. Section 8.3.5 contains the current solution to this problem.

The literature (e.g. Aron and Muellbauer [1], Viglione [32]) discusses reasons why published fatality counts may be much smaller than the real death toll, and this works by considering *excess mortality* (Übersterblichkeit). There is a database [30] kept by *The Economist* with fairly recent values, see Table 7 as retrieved in mid-September, using data up to August 28th. The figures are based on EUROMOMO[6], a European mortality monitoring activity. Their COVID-19 figures may not be the same as the Johns Hopkins figures, but the factors between similar data acquisition strategies for mortalities might be portable. They are now in use for the cumulative deaths D when importing the Johns Hopkins data. They might be time-varying as well, but we ignore this, so far.

10 Changes in Programming

10.1 Technical Notes

The original programs for the published paper are frozen, and available via http://num.math.uni-goettingen.de/schaback/research/papers/OC19M.zip from the author's research website. But there are certain changes made in the meantime that will be reported here. New software versions ara available on request, and the programs producing the plots for the September 2020 version will be frozen as well.

[6] https://www.euromomo.eu/

Country	Factor
US	1.41
Spain	1.55
Italy	1.25
Sweden	0.92
France	0.97
Germany	1.01
UK	1.13
Austria	2.32
Denmark	0.79

Table 7 Factors for changing COVID-19 death data by results of excess mortality.

10.1.1 Useful Data

Basic Johns Hopkins data plots will still show all data. But for any prediction algorithm, outbreak data are ignored before they reach at least 10 deaths and 100 Confirmed. This rule complies well with Figure 4. This has a serious consequence for modelling re-infections. As stated above, one has to wait for 10 new deaths or 100 new Confirmed to get useful data for a re-start of the models. So far, test runs with full new starts are not provided. The main reason is to prefer longer calibration periods than those for a late restart.

Section 8.4.3 provides a method for eliminating non-monotonicity errors and transmission delays, coupled with a smoothing technique. This was applied for predictions and estimates from July 2020 on, and revised for the September version.

10.1.2 Estimation Algorithm for Case Fatalities

The optimizations in (28) and in (2) of [25] can be simplified by taking the q_i as variables under a nonnegativity constraint. This allows to use linear least squares routines. The results are the same except for roundoff, for all cases seen, and the solution now is unique. The interpretation of the p_j as probabilities are lost, and the q_i describe nonnegative portions of the newly Confirmed. The constraint on the sum of the q_i turned out to be automatically satisfied in all test runs performed so far, but it was added into the new software version. As a side effect, it shortened runtime.

After July 2020, the algorithm was run under the additional linear constraint $I = C - R - D \geq 0$. This constraint was active only for countries where the D values were exceptionally polluted. After the changes in the technique of section 8.4.3 were made in August/September, the constraint was not violated anymore on the examples presented, but it still is necessary for Austria and Switzerland.

10.1.3 Replacing Constants by Time Series

As mentioned before as a change to the models, the case fatality rate γ_{CF} is now calculated as a time series via section 4.3.2, not globally. The k-day rule is applied for $k = 14$ using data for $2k$ days backwards. The infection fatality then enters as a time series (43), but both changes only concern the calibration phase of the full model.

In particular, the variation of the detection rate now leads to different and hopefully more realistic estimations of the hidden Infectious M and the hidden Recovered H.

Experiments indicate that also the number k for the k-day rule should be time- and country-dependent, but this is currently not implemented. Increased experience with COVID-19 treatment of severe cases seem to have led to longer hospitalization and lower death rates, as soon as health systems were not overloaded.

10.1.4 Starting Values for the Calibration

Already in section 4.5.1 of the basic paper, the starting values for the calibration phase were no big problem, because the observed C values dominate what happens. See Figure 10 for a demonstration. The effect is even stronger now, because the calibration phase is longer. The new software, however, uses a different strategy that turned out to start closer to what happens shortly after startup. The idea is to take the ratios $I/(I+R)$ and $R/(I+R) = 1 - I/(I+R)$ of the observed Johns Hopkins data, using a 7-day backlog. These ratios should roughly be the same as those for M and H of the hidden model. Therefore the M and H starting values for calibration are now set by splitting $N - S - C$ using these ratios. This maintains the balance $N = S + H + M + C$. The difference to the old rule has no long-term effect, i.e. on the outcome of the calibration needed to start a prediction.