Model-based and model-free characterization of epidemic outbreaks — Technical notes on Dehning et al., Science, 2020

Jonas Dehning[1], F. Paul Spitzner[1], Matthias Linden[2], Sebastian B. Mohr[1,3], Joao Pinheiro Neto[1], Johannes Zierenberg[1], Michael Wibral[4], Michael Wilczek[1,5], and Viola Priesemann[1,5,6]

[1] *Max Planck Institute for Dynamics and Self-Organization, Am Fassberg 17, 37077 Göttingen.*
[2] *Institute for Theoretical Physics, Leibniz University, 30167 Hannover, Germany.*
[3] *Institute for Theoretical Physics, University Leipzig, Postfach 100 920, 04009 Leipzig, Germany.*
[4] *Campus Institute for Dynamics of Biological Networks, University of Göttingen, Hermann-Rein-Straße 3, 37075 Göttingen, Germany.*
[5] *Institute for the Dynamics of Complex Systems, University of Göttingen,Friedrich-Hund-Platz 1, 37077 Göttingen, Germany.*
[6] *Bernstein Center for Computational Neuroscience, Hermann-Rein-Str. 3, 37075 Göttingen, Germany.*

(Dated: June 8, 2020)

In this technical note, we provide additional background on our Bayesian inference for change-point detection in COVID-19 case numbers (Dehning et al., Science, 2020). In particular, we explore basic properties of model-based and model-free estimates of the reproduction number, discuss what conclusions can be drawn from Bayesian analyses, further develop our model and apply it to newly available data, and discuss potential issues with changes in testing policies.
This technical note presents work in progress and should be considered like an internal draft. It is not ready for submission yet, and is being frequently updated.

CONTENTS

I. INTRODUCTION

After the initial release of our manuscript "Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions" in Science [1], we have received many constructive comments and interesting questions, and have also faced some recurring misunderstandings. This technical note is intended to answer the most important of these questions, to give additional background for understanding our results, and to also discuss the robustness and performance of our model in the light of newly available data, in particular data based on on symptom onset times.

The inspiration and comments we received can be broadly categorized into four topics:

1. Remarks on apparent discrepancies between the values for the reproductive number R as reported by the Robert Koch Institute (RKI) and the corresponding spreading rate resulting from our published analysis. We will explain below how this

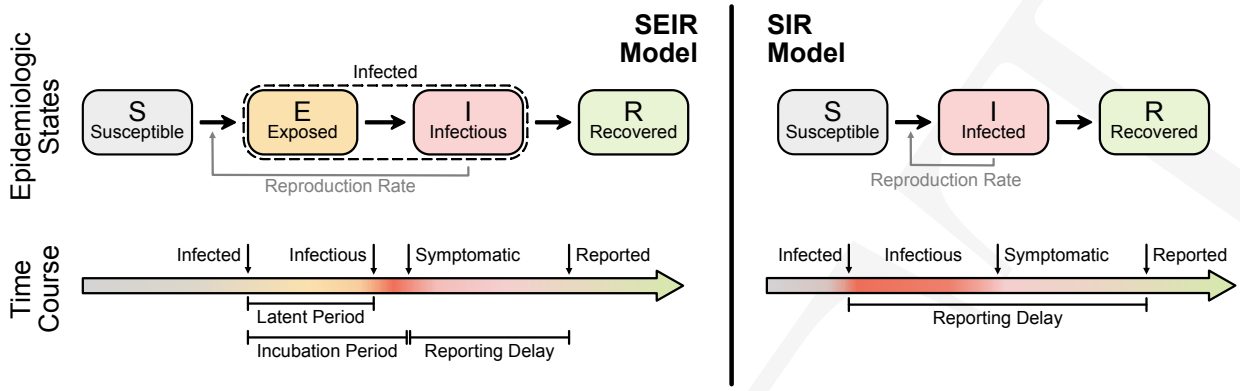

FIG. 1. Illustration of two basic compartmental models in epidemiology. The SEIR model (left) captures the basic steps that infections passes through: A healthy person becomes infected (leaves S, enters E) but not infectious; after some time ('latent period') the person becomes infectious (leaves E, enters I) but symptoms only show after some incubation period; after some time the person is no longer infectious (leaves I, enters R), which can have several reasons including isolation, conventional recovery, or death. The SIR model (right) is the most basic compartmental model and does not distinguish between infectious and infected: A healthy person becomes infected (leaves S, enters I) and by this begins to infect other persons, but only shows symptoms with some delay; after some time the person "recovers" (leaves I, enters R), which again includes isolation, recovery, or death.

apparent discrepancy arises from the comparison of model-free estimates to those from a differential-equation based modeling of disease dynamics. We show how the model-free approach may substantially underestimate the reproductive number R immediately after a sudden drop in R has occurred. From the comments we received it seems that this very important fact related to estimating R is largely unknown, and also counterintutive to most readers. This effect, however, fully explains the apparent discrepancies between the RKI reports and our study. We therefore derive and demonstrate it in detail here.

2. Questions revolving around the philosophy and interpretation of our modeling approach that combines a differential equation model of the disease outbreak, Bayesian parameter inference and Bayesian model comparison. Most frequently we were asked if and in what sense our results have a causal interpretation. As we will explain below, our approach selects the most plausible of multiple causal explanations of the observed data, but does not establish strict interventional causality.

3. New data have been released in the time since our analyses were completed. Most prominently, data on the exact times of symptom onsets (epi curve) are now available and supersede the case report data as the best data source for modeling the outbreak. As we will show below, our conclusions remain unchanged when updating our model to the new data.

4. Questions on how changes in testing capacity may have influenced our results. Given the data that have become available on the weekly (daily) number of performed tests, test capacity, and on delays between symptom onset, test and case report, we reanalyze in great detail the disease and testing dynamics, especially with respect to the timing of the peak in new symptom onsets. We conclude that all symptom onsets that are relevant for the main conclusions of our previous publication have been tested at a time when testing had sufficient capacity and was sufficiently constant.

We will in the following address the issues revolving around the reproductive number R first, also introducing the basic terminology of disease spreading and the fundamental difference between model-free and model-based estimation of epidemiological parameters. Next, we will discuss philosophy and interpretation of model-based estimation in the Bayesian framework and the causality question. We then show how our original analyses can be evolved to incorporate new data, in particular on symptom onset (epi curve). Last we turn to the important question of testing.

II. ESTIMATING THE REPRODUCTIVE NUMBER

A. Basic SIR dynamics

Before we define the reproductive number R, we briefly recapitulate the basic SIR dynamics that we consider (Fig. 1). In principle, the course on an infection can be described as follows: A susceptible person (not infected and not immune) becomes infected but is initially not infectious; after some time, the person starts to be infectious but symptoms only show after the incubation

period; eventually, the person is no longer infectious because it is either isolated, it recovers, or it dies. The idea of compartmental models is to group the population into compartments; in the most simple but established SIR model these are susceptible (S), infected (I), and recovered (R). Assuming a well-mixed population (a mean-field approximation of everybody interacting with everybody), one can formulate differential equations that describe the time development of these compartments:

$$\frac{dS}{dt} = -\lambda \frac{SI}{N} \tag{1}$$

$$\frac{dI}{dt} = \lambda \frac{SI}{N} - \mu I \tag{2}$$

$$\frac{dR}{dt} = \mu I \tag{3}$$

This assumes a spreading rate λ for infected people to infect susceptible people (who they meet randomly) and a recovery rate μ for infected people to recover. These differential equations can be extended to include various different compartments, in order to better resolve the temporal course of the disease, but typically keep the mean-field assumption of a well-mixed population unless evaluated on some (typically unknown) network. In this case, additional compartments reflect spatial information.

Observed case numbers are always delayed from the true infection date (Fig. 2). In general, when a person becomes infected, the onset of symptoms is delayed by the incubation period. Upon symptom onset, it typically takes a few days until the person undergoes a test and the case is reported (although some people are tested before symptom onset, e.g. if contacts are traced or tests are performed at random "Stichprobe"). However, for the modeling, one is usually interested in the actual time when a person becomes infected — but this information is not directly available in real-world data. One either works with the reporting date or with the dates of the symptom onset (epi curve) that can be reconstruced e.g. via nowcasting. Note that these are still delayed with respect to the true infection dates due to the incubation period. For the example models in the following, we synthetically generate observed cases — symptomatic or reported — by convolving the infected cases with a distribution of incubation periods or reporting delays, respectively (Fig. 2).

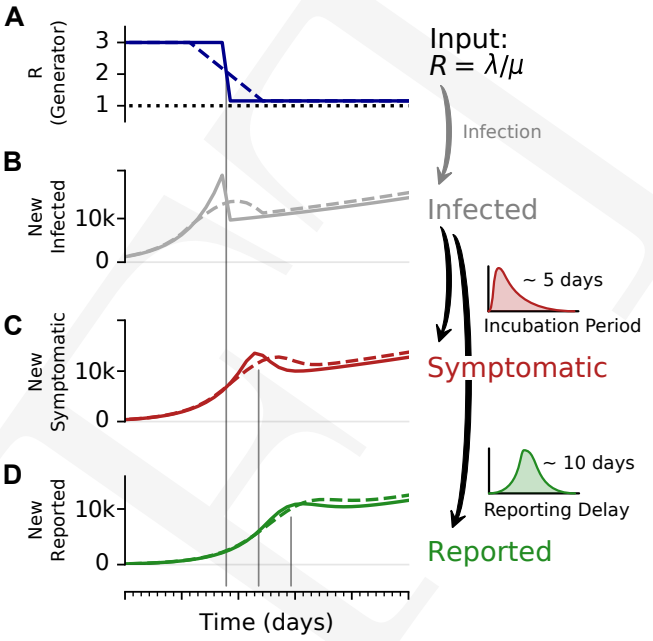

FIG. 2. **A change-point in R can lead to a transient decay in case numbers.** To illustrate the effect of a change point, and the delays in observing symptomatic and reported cases, we consider an SIR model with a fast or slow decay of R, and generate synthetic case numbers. **A** The reproductive number R exhibits a change point from $R = 3$ to $R = 1.15$, with a duration of either 1 day (solid) or 9 days (dashed). **B** The number of new infections can show a transient decrease caused by the change point in R, even though the underlying dynamics are always in the exponentially growing regime of $R > 1$. Such a decrease can be misinterpreted as $R < 1$. The number of **C** new symptomatic cases, and **D** reported cases is generated by convolving the new infected with a log-normal incubation period (median = [XX] days) or reporting delay (median [xx] days), respectively. Note that the convolution shifts and smooths the curve of the new infected. Nonetheless, the counter-intuitive effects of a transient decrease in case numbers caused by a change points, is still very well visible (See Fig. **??** for the challenges of estimate R in around the change point.)

B. Model-free estimation of reproduction number R_t

Definition of R. The reproductive number R quantifies how many susceptible persons are on average infected by one infected person. If one person infects on average more than one ($R > 1$), then case numbers are growing exponentially. If in contrast one person infects less than one ($R < 1$), then case numbers are declining. Therefore, $R = 1$ marks the critical stransition between growth and decline of case numbers. Estimating the reproductive number R in principle can be done in two manners, either by inferring it from observed case numbers, or by following infection chains step by step. If one infers it from observed case numbers, there are a number of possible approaches. Some approaches are summarized in Fig. 4 and detailed below. All these approaches are applied to the observed case numbers (day of symptom onset, i.e., epi curve, or day of reporting). In the following we assume that they are applied to the epi curve.

The most straight-forward definition of the reproductive number assumes a reproductive process with offspring generation, such as a branching process [2]. For this, one

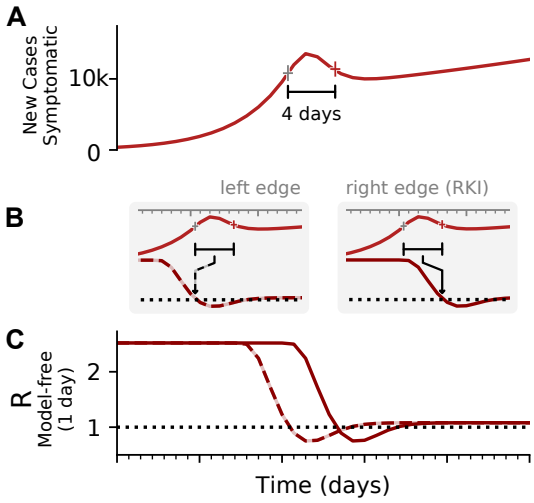

FIG. 3. Two different conventions to define the reproductive number R: Infections in the future or infections from the past. **A** Synthetic data for new symptomatic cases. The marked interval indicates an assumed generation time of 4 days. **B** The basic reproductive number can be defined either on the left edge of the generation interval (left, dashed line), describing the average number of future infections that are cause at time t, or on the right edge of the interval (right, solid line), describing the average number of infections at time t that were caused by the past ones. **C** Depending on the convention, the resulting curve of R is shifted by the generation time g. Note that in both cases the R is estimated erroneously to fall below $R = 1$, although in the underlying model it was was $R > 1$ all the time. This is an effect of the SIR dynamics together with a change point in the underlying R. (See Fig. 4 for model details.

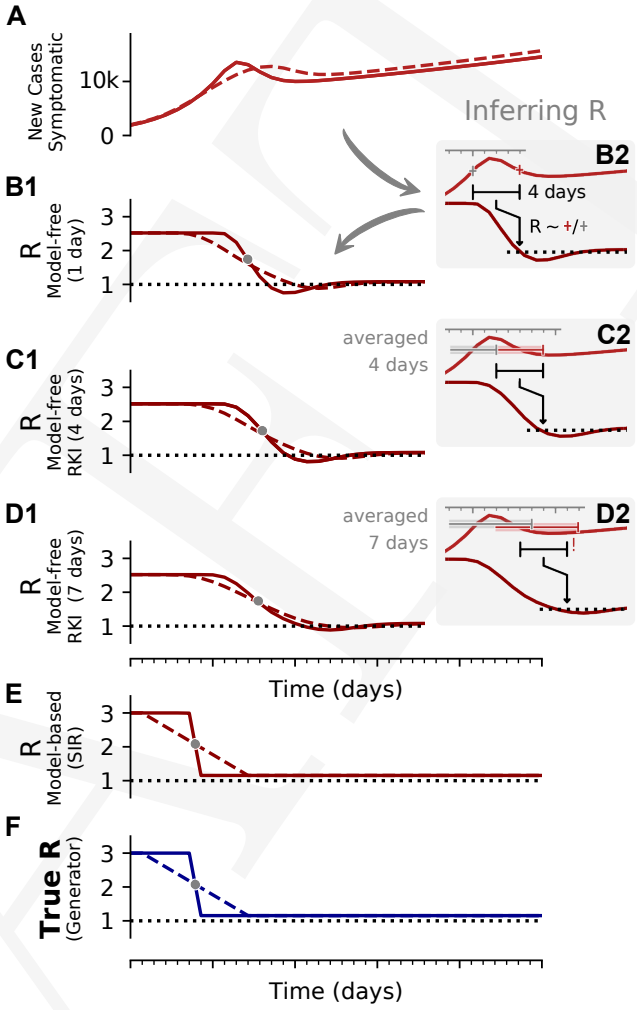

FIG. 4. The inferred reproductive number R depends on the inference method. **A** Synthetic data for new symptomatic cases generated with SIR dynamics from an underlying R with one change point (see **F**) of duration 1 day (solid) or 9 days (dashed). (See Fig. 2 for details). **B** Model-free inference of R based on the ratio of case numbers at time t and time $t - d$, marked by a red and gray cross (inset), respectively ('right-edge convention', cf. Fig. 3). **C** Model-free inference of R following the Robert Koch Insitut convention, i.e. using the definition of **B** but with averaging over a window of the past 4 days (inset, red and gray bars). **D** Same as **C** but averaging over 7 days. Note the overlap of intervals. - All the model-free methods (**B**-**D**) can show an erroneous estimate of $R < 1$ transiently, due to the change point in the underlying true R (depicted in **F**). **E** The inferred R using change-point detection with an underlying dynamic model (SIR) does *not* show a transient erroneous $R < 1$ period. If the underlying dynamic model corresponds well enough to the true disease dynamics, then this approach reproduces the true R (**F**) that was used to generate the data (**A**).

assumes a generation time g in which an infectious person can generate offspring infections. In the simplest case, one could consider that offspring infections occur exactly after one generation time g. This allows to infer the effective spreading rate R precisely:

$$\hat{R}_t = \frac{\text{number of newly infected at time } t+g}{\text{number of newly infected at time } t} \tag{4}$$

$$= \frac{C_{t+g}}{C_t}. \tag{5}$$

In reality, these newly infected case numbers C_t have to be approximated, e.g., by using new symptomatic cases or new reported cases. Moreover, the generation times g of each infection are widely distributed, so that using the average value g (or an estimate of it) is a further approximation. For its simplicity, this inference of R is widely applied and has proven quite useful.

When go into detail, there are two different conventions for the timing of the estimated reproductive number with respect to the case numbers (Fig. 3). Above, we consider R_t to characterize the number of future infections that are caused by infections at time t (left-edge convention). Alternatively, one can consider R_t to characterize the

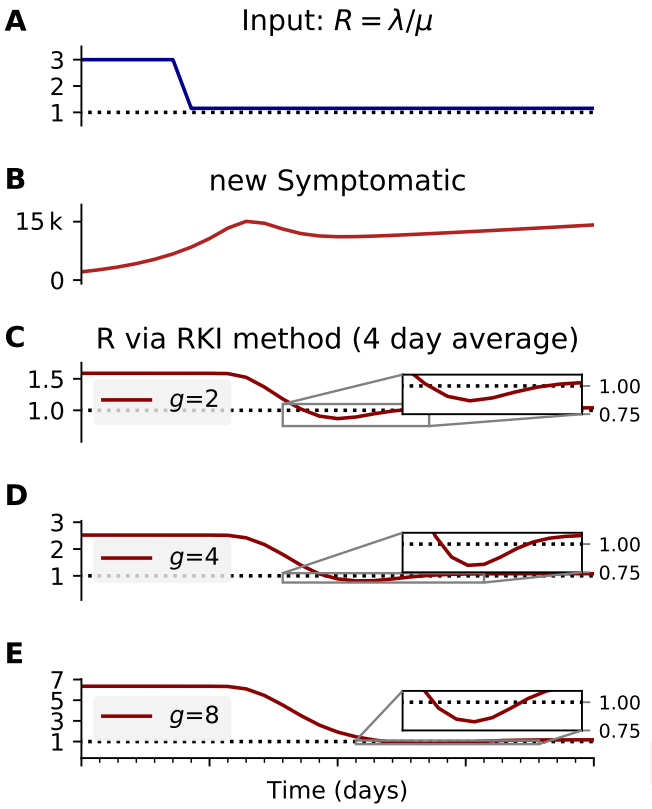

FIG. 5. Inferred reproductive number depends on the assumed generation time g. We generate synthetic data using SIR dynamics with time-dependent R including a 1-day change point (**A**, cf. Fig. 4) that yields new symptomatic cases with transient decrease (**B**) despite all $R > 1$. Using the RKI convention to infer R (4 day average, right-edge convention), we demonstrate how generation times (g) result in different R curves (**C**-**E**). In particular, we find different initial levels of R (left plateau), differently long crossover duration (time from left plateau to right plateau), and differently deep transients of $R < 1$ (see insets).

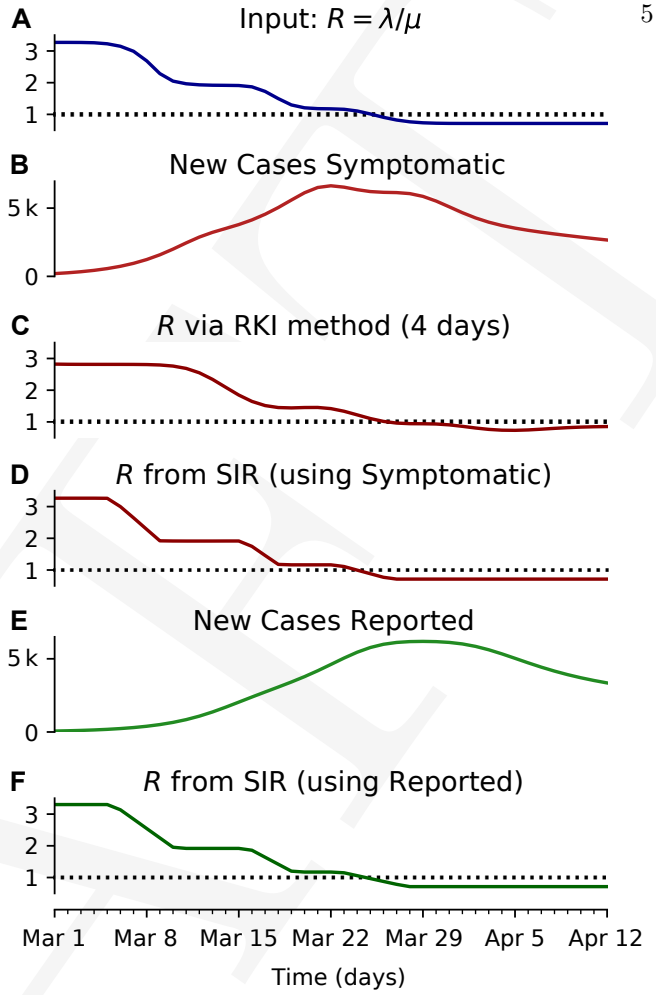

FIG. 6. The change-point detection methodology yields consistent results irrespective of whether it is applied to the new reported cases or the new symptomatic cases (e.g. obtained by nowcasting). **A** Time-dependent reproductive number as inferred from case numbers in Germany [1]. **B** Synthetic data for new symptomatic cases generated with SIR dynamics from the underlying time-dependent R (see **A**). **C** Inferred R from new symptomatic cases using RKI method (4 days generation time, right-edge convention) would reproduce step-like behavior (no noise present) but prematurely drops below $R = 1$ (dashed line). **D** Inferred R from new symptomatic cases using change-point detection with dynamic model (SIR) correctly reproduces the input (**A**). **E** Synthetic data for new reported cases generate with SIR dynamics as in **B** (cf. Fig. 2). **F** Inferred R from new reported cases (**E**) using change-point detection with dynamic model (SIR) also correctly reproduces the input (**A**). Note that both **D** and **F** show sharper steps because of the assumed piece-wise linear change points in the model, and that they perform so well because they employ the true dynamic model that is used for the synthetic data. Both are model assumptions that need to be justified in our approach.

number of infections at time t that were caused by the past pool of infected (right-edge convention), defined as

$$\hat{R}_t = \frac{\text{number of newly infected at time } t}{\text{number of newly infected at time } t-g} = \frac{C_t}{C_{t-g}} \tag{6}$$

The results for R are exactly equivalent, apart from a shift in time by exactly g.

R as calculated by the RKI. Real-world data are often noisy, and therefore averaging over a certain time window can help to smooth the estimate. This procedure is used in two variants by the RKI, smoothing over four days or over seven days[3]. In both cases, they assume a constant serial interval (generation time) of $g = 4$ days (Fig. 4). The four-day smoothing has the advantage that it reacts a bit faster, the seven-day smoothing has the advantage that it smooths out weekend-related modulations of test numbers.

The general equation then reads as follows:

$$R_t = \frac{\sum_{j=t-w}^{t} C_j}{\sum_{k=t-g-w}^{t-g} C_k}$$

w is the window. The Robert Koch Institute chooses a window of 4 or 7 days [REF].

C. Model-free methods versus model-based methods to infer reproductive number.

In order to demonstrate potential issues when inferring the reproductive number R, we systematically compare model-free methods and model-based methods on synthetic data from an SIR model (Fig. 2). With model-free methods, we refer to inference methods for R, which do not explicitly incorporate disease dynamics (SIR). The three methods we presented above belong to this group. These methods to estimate R are straight forward and easy to implement. However, they might lead to biased estimates when R is changing rapidly. More precisely, in the following we show that these methods (1.) smooth out fast changes in R, (2.) produce some delay compared to the underlying R, (3.) the estimate depends on the assumed generation time, and (4.) around change points they may return transiently $R < 1$, even if the true value was never smaller than 1. While these methods are very useful for a fast estimate of R when R is not changing too quickly, they may lead to wrong estimates otherwise.

1. *Model-free methods may smooth out fast changes.*

In Fig. 4, the $\hat{R}$ that is inferred by model-free methods undergoes a smoother change than the true R. The smoothing has two origins: First, when using the sliding-window of four or seven days (RKI methods), multiple days are combined to obtain an R value for one day. Second, R has to be calculated from the daily new symptomatic or reported cases (Fig. 2 C, D), because the dates of infection (Fig. 2 B) are not directly accessible in real-world data. As discussed before, symptom onset and reporting date are delayed from the infection date. Because the delays vary from case-to-case, these two curves are smoothed out compared to the infection curve (In other words, the smoothing originates from the variance in incubation period and reporting delay, see later Fig. 10 in the section about testing). Hence, if smoothing is not explicitly incorporated in the inference of R, fast changes appear slower than they truly are, and successive fast changes may appear as a long transient.

2. *Model-free methods produce delayed estimates that are difficult to interpret*

In our example in Fig. 4, we estimated $\hat{R}$ based on the number of new symptomatic cases as produces by our model. The $\hat{R}$ of all three model-free methods is shifted in time compared to the true R (Fig. 4F).

How does one interpret the shift and where does it come from? To interpret the shift and compare between the different methods, we focus on the time point where half of the steep step in R has been detected (gray dots). This shift has multiple contributions. One contribution originates from using the dates of symptom onset, which is shifted on average by the incubation period (in our example ≈5 days). This generates the 4-5 day shift of the one-day method (Fig. 4B). Because the incubation period is not constant and typically is asymmetric, there is an additional asymmetric distortion towards either direction, depending on the shape of the actual distribution of incubation periods. Another source for the shift comes from the time average, which explains the additional (approximate) 1-2 day shift in the four-day and seven-day methods employed by the RKI (Fig. 4C,D). Because of the specific definition of the position of the 4 and 7-day window of the RKI, the two versions of $\hat{R}$ have a very similar average delay of 5-6 days in total with respect to the true R.

Both, the variable incubation time and the time averaging also impact the start- and end-points of the change in a non-trivial manner. In combination, multiple sources cause shifts that can point into opposite directions. While the sources can be identified conceptually, the combined effect cannot perfectly be disentangled or compensated.

Due to multiple sources of shifts and smoothing, a simple post-hoc shift of the R-curve cannot reproduce the true R around a change point. For example, a shift of Fig. 4D by 5 days would suggest a start of the change point before it starts in reality (Fig. 4F). This fact has led to multiple prominent misunderstandings in relation to the RKI data and the effects of governmental interventions. Instead of shifting curves to partially correct for one or another potential delay, an inference of R using model-based methods can account for this and other potential biases. When using a good model, such a model-based approach returns the correct R with the correct steepness and time point (Fig. 4E).

3. *R estimates depend on the assumed generation time.*

The assumed generation time g impacts the absolute value of the estimated reproductive number R (Fig. 5). We exemplify this effect using the method of the RKI (4 day average), where we vary the assumed generation time g. (Note that the same effect applies to model-based inference.) In a stationary phase with constant R, the case numbers change by a factor R within one generation. Within two generations, they change by the factor R^2, and so on. Hence, when assuming erroneously double (or half) generation time, then one obtains the square (or square root) of the true R as estimate. More generally, $\hat{R} = R^{g_{\text{assumed}}/g_{\text{true}}}$.

In the example, we assumed three different generation times (2, 4, and 8). At the onset, $\hat{R} = \simeq 1.6$, $\hat{R} \simeq 2.56$, and$\hat{R} \simeq 6.5$ for $g = 8$, as expected from theory. In absolute terms, this dependence is less pronounced near a reproductive number of 1; assuming e.g. $R = 1.1$ then

assuming double (or half) the generation time results in $R = 1.21$ (or $R = 1.05$). - This small example shows that estimating the absolute value of the reproductive number from observed case numbers without knowing the precise generation time may lead to misestimates.

4. *Model-free methods may return erroneous transient periods of $R < 1$ at change points.*

In our examples (Figs. 4 and 5), we consider that R changes rapidly from $R_0 = 3$ to $R_1 = 1.15$ within one day. Such a sudden change leads to a transient decrease in new case numbers — despite $R > 1$ always. How can there be decrease in new cases although $R > 1$? The transient decrease results from the pool of infected suddenly infecting considerably less people. This decrease in infections causes the sharp peak and a sudden drop in new infections (Fig. 2B, solid line). It then carries over to the number of new symptomatic and new reported cases, with the respective delay and smoothing (Fig. 2C,D]). This transient decrease depends on the duration of the change point: While it is strongest for steep changes, it also occurs for a nine-day change point (Fig. 2, dashed line).

Naively, a transient decrease might be interpreted as a transient $R < 1$, but that is not the case here. A model-free method cannot distinguish between different causes for transient decreases in case numbers, being it due to transient non-linear effects (Fig. 2) or due to a true exponential decay ($R < 1$). The model-free methods in our example (Figs. 4 and 5) correspondingly yield non-negligible periods of $R < 1$, even though the underlying model dynamics have $R > 1$ always. Model-based approaches, on the other hand, can account for transient non-linear effects if included in the model, e.g., as change points, and — if the model is correct — even reproduce the true underlying dynamics (Fig. 4E). To conclude, if one infers R in a model-free manner, by computing ratios of case numbers, then the local minimum leads to an erroneous estimate of $\hat{R} < 1$ (Fig. 4B,C.D).

5. *Well chosen model-based methods can reconstruct complex disease dynamics*

When the chosen model describes the true disease dynamics well, robust inference of the true underlying reproduction number (and other parameters) is possible. To demonstrate the robustness of model-based inference, we generate synthetic data using an SIR-model as inferred from case numbers in Germany between March 2 and April 21 [1] (Fig. 6). The Bayesian model inference can recover the reproductive rate (Fig. 6D,F), whereas with the model-free method, the recovered R is slightly biased (Fig. 6C). Note, however, that the model has to match at least approximately the disease dynamics, to allow a good inference. This is why we used different models to assess the robustness of our results in Ref. [1] (SIR: Fig. 3, SEIR-like: Fig. S3, SIR without weekend modulation: Fig. S4).

III. WHAT CONCLUSIONS CAN ONE DRAW FROM A BAYESIAN ANALYSIS?

A. Modeling background

When the Coronavirus-pandemic arrived in Germany we set out to model the spread of the disease as rapidly as possible. Thus, our model from the start was aimed at giving estimates with their corresponding error bounds based on the data available at that time. To this end we decided to use a Bayesian strategy as it allowed formulating well-documented assumptions on those aspects not available from data at that time. Within the Bayesian framework these assumptions can and should be replaced by data as soon as these become available, and we implement such an improvement below for the case of information on symptom onset times that have become available in the meantime. Given such new data it will also be interesting to evaluate post-hoc the assumptions and the performance of our model . This will also give some guidance as to whether to employ a model of this kind again in a new scenario (another disease outbreak or pandemic) where some relevant data will also not be available immediately. We note that taking these steps is the intended development in Bayesian inference.

We also note that all statistical procedures come with their own assumptions, e.g. on distribution of the data, models of measurements and random errors. Bayesian analysis is no exception to this rule; in our view the only difference is that modeling assumptions are not taken for granted based on the long-established used of a method (say, a t-test) but need to be formulated anew for each case. The fact that the assumptions are hand-tailored to the application case may seem subjective sometimes; yet, similar assumptions are being made, more tacitly perhaps, in other frameworks, as well. This said, it is nevertheless important to question and discuss (our) modeling assumptions and to test the sensitivity of our results to the modeling assumptions. As far as space restrictions allowed we have discussed our assumptions already in the main manuscript [1], but we here give a much deeper and broader and more educational treatment.

B. Bayesian inference as reasoning under uncertainty, bound to be updated

The results of a Bayesian analysis at some publication time point T represent what we should believe in at that time point T, given the knowledge available at T (causes and data known at T). These results represent something that we should be able to agree on given the knowledge at T (and some practical constraints, see below), but these

results may change given more information at a later time $T + \Delta_T$. Changing ones mind with the availability of additional information is designed into Bayesian inference as "the logic of science" (E.T. Jaynes) from the start. In other words, scientific inference and the associated models are bound to be updated - just like the relativity theory and quantum theory in physics overrode their former model counterparts. The important question is thus not whether a model is correct in absolute terms, but whether it was possible to agree on the model (and the inference provided by it) at time T, and also if the inference provided at T was robust, for example in the sense that the credible intervals for the model parameters at T comprise those obtained at $T + \Delta_T$.

From this perspective it is obvious that now, more than a month after finalization of our published analyses on April 21st, new data have become available and that the model can, and should, be improved accordingly. Important data in this respect are the reliable data on putative infection dates which at present take about 7 days to come in for at last 80% of the cases (Fig. 10), and which where only published more recently than our internal analysis cut-off. We present results obtained using these data below and compare them to our published results.

C. Conditions for plausible alternative models entering model comparison

A frequent, and important misunderstanding around Bayesian model comparison is that one is allowed to formulate very many models at random and then let the data decide on the best model via the Bayesian model evidence (or the LOO-scores). This notion fails to notice that the model evidence $p(D|M_i)$ is only one part of the decision on the preferred model. The formal equation for deciding between models i and j would be:

$$\frac{p(M_i|D)}{p(M_j|D)} = \frac{p(D|M_i)}{p(D|M_j)} \frac{p(M_i)}{p(M_j)} , \tag{7}$$

i.e. taking such a decision entails accounting for a-priori plausibility of different models, i.e. $p(M_i)$ and $p(M_j)$. While it is customary to assign equal a-priori plausibility to all the models being considered, this does not mean that just any model qualifies for use in this decision procedure. Rather, each model subjected to a model comparison needs to be well justified. This is one of the reasons why we did not consider for example models of sustained, constant drifts in the effective spreading rate λ^* (or, equivalently the reproductive number R), as we did not come up with plausible explanations for such a behaviour (except perhaps arguments based on herd-immunity, which seem implausible now, in the light of second waves of infections and a recent rise in λ^* from its all-time low, and also in the light of country to country comparisons, Fig. 7).

On a practical note, useful modeling also has to reflect certain limits on model complexity in relation to the available data, and also computational resources. Known phenomena, that can nevertheless not be modeled must therefore often be integrated into noise terms that are designed accordingly (as was done with the modeling of observation noise in our case, instead of using full stochastic differential equations). The best that can be done then is to investigate the sensitivity of results with respect to the simplifying assumptions that have been made.

It is also in order to explain in simple terms how results of a Bayesian analysis may be interpreted: In the Bayesian framework probabilities are measures of the plausibility of statements about the world, given our present knowledge. Thus, the results of Bayesian parameter inference for example indicate credible (plausible) ranges in which we should assume the unknown parameters to be. Assuming them to be elsewhere with high probability would be inconsistent with the information we have. In this sense, these credible intervals may form the basis for decisions we have to take.

D. Models as competing causal explanations of data

Last, we note that the notion of causality resides only in the construction of the models – with different models incorporating different possible causal explanations (e.g. in the form of differential equations for the disease dynamics) of the data. Performing model comparisons then selects more plausible over less plausible explanations, but does not provide a proof of causality in the strict sense advocated for example by Judea Pearl [4] or by Ay and Polani [5]. Yet, fulfilling the formal criteria for causality in this strict sense would need multiple replications of the pandemic process, each time with different settings of the relevant variables, such as interventions. Even when treating the SARS-CoV-2 outbreaks in different countries around the world, with their different interventions (or lack thereof), as replications establishiung formal casuality may remain an elusive goal due to multiple other variations from country to country. In sum, the results of our Bayesian analysis must be seen as a search for the most plausible causal model of the data, given the data available at the time of analysis, and as providing credible ranges of the parameter values relative to this most plausible model.

Later, discussions (such as the one presented here) of the selected models and the inferred parameter ranges should then investigate and update modeling assumptions, and reason whether the causal model can be maintained, or not.

When analyzing improved data that reflect the dates of symptom onset rather than case reports to improve our modeling we find that both the preference for a three change point model as well as the inferred parameter ranges do not change drastically, and we maintain our original interpretation of the pandemic process and the effectiveness of governmental interventions.

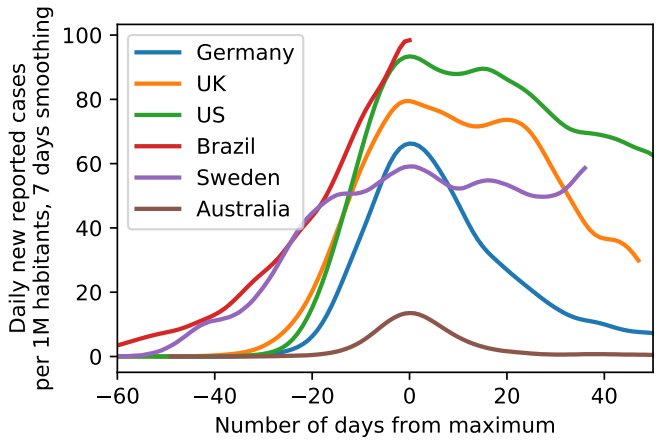

FIG. 7. Comparison of the case numbers per one million inhabitants of exemplary countries as illustration of the range of possible case numbers developments. Note how both the peak height as well as peak width of some countries are considerably larger than for Germany, providing evidence against saturation effects ('herd immunity') in Germany (Data until June 3, 2020).

Last, alternative models assuming herd immunity as a reason for the sustained observed drop in infection rates still do not seem plausible to us in the light of rapidly surging second waves or sustained high levels of new infections (such as in Sweden, see Figure 7).

IV. MODEL EVOLUTION

Modeling efforts at the beginning of an epidemic outbreak are aimed at providing a rough but timely and robust description of the disease outbreak, making use of whichever data are available at that time. Later modeling efforts in contrast make use of more detailed data and provide deeper insights into how the outbreak unfolded. While these latter models are useful for a better understanding after the fact, they cannot be applied early on due to a lack of data, and often cannot inform decisions fast enough. However, a comparison of early and later models can provide important insights about the robustness and usefulness of the early models with respect to the later ones (here usefulness means that the early models describe the epidemiological parameters and their uncertainties well enough to inform decisions). For the case of the COVD-19 outbreak in Germany, the initially available data were sorted based on date of reporting, where the reporting occurred after an unknown delay between symptom onset and report. Only later, data organized by time of symptom onset, the so-called epi curve, became available. Even after their initial release, these data were still updated and refined (see Fig. 8); also note that data for symptom onsets still take some time to arrive and be compiled, i.e. the delay between symptom onset and testing/reporting is still considerable (see Fig. 13). In particular, this means that reliable epi curve data for April 21st, our analysis cut-off date in [1], were not available until much later. Now that these data are available, however, we can compare models based on data organized by reporting date, modeling the reporting delay and incubation period, and models based on the epi curve, modeling the incubation period only.

A. Model updates based on time of symptom onset and comparison to previous results based on time of reporting

Ideally modeling of an epidemic outbreak should rely on data organized by infection date - yet, such data are rarely available outside of the analysis of individual, well-confined infection chains. The next best option then are data organized by date of symptom onset - the epi curve. Naturally, symptom onset precedes the test and report in time. Thus, the epi curve is only available after a certain delay, which can be substantial. Furthermore, the time of symptom onset may remain unknown for a significant fraction of reported cases. If so, then reconstructing the epi curve requires data imputation and further modeling (e.g. nowcasting [6, 7]), which may further delay the availability of this curve. At the initial stages of an outbreak one may therefore decide to analyze data organized by reporting data. For a comparison of analyses it is important to understand how the curve of reporting dates and the epi curve are linked. Both curves originate from the curve of initial infections by a convolution (see Fig. 2). The epi curve is the curve of initial infections convolved by the distribution of incubation periods, while the curve based on reporting date is the curve of true infections convolved by the (less well known) distribution of delays between infection data and reporting date. Technically, a report can happen before symptom onset, albeit this may be rare. Therefore, the curve of reporting dates is not exactly a convolution of the epi curve with an additional delay distribution.

We have reanalyzed the initial stages of the outbreak until April 21st based on the epi curve that has become available (see Figs. 17 and 19), using models with one, two and three change points, based both on SIR and SEIR dynamics (only figures for the three change points models are shown).

These new results do not change our main conclusions presented in [1]. Specifically, model comparison still favors the three change point models over their simpler counterparts (see table I), and only the third change point leads to a value of the spreading rate λ^* that is clearly below zero. At the quantitative level, however, we see some evidence for a larger drop introduced by the first change point when using the epi curve data, and smaller drops induced by the second and third change point, especially when using an SEIR model (see Fig. 19). These quantitative changes are driven by the epi curve dropping faster than the curve reflecting reporting date (see Fig. 9C).

TABLE I. Model comparison: Using leave-one-out (LOO) cross-validation, we compare the SIR and SEIR model variants using the epi curve as data (Figs. 17 and 19). Lower LOO-scores represent a better match between model and data (pLOO is the effective number of parameters).

Model	# c-pts.	LOO-score	pLOO
SIR main	0	900 ± 13	6.36
SIR main	1	774 ± 14	12.72
SIR main	2	755 ± 13	12.17
SIR main	3	725 ± 15	19.66
SEIR-like	0	900 ± 14	6.65
SEIR-like	1	749 ± 12	8.05
SEIR-like	2	739 ± 13	10.28
SEIR-like	3	726 ± 14	14.04

In sum, we conclude that the original model based on data organized by reporting date was useful to understand disease dynamics in the absence of the epi curve and robust in the sense that its main results still hold.

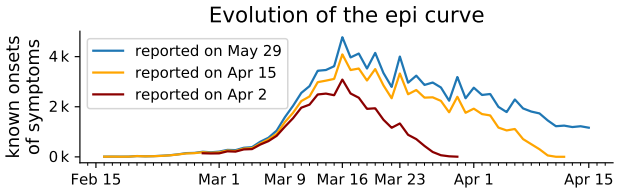

FIG. 8. The numbers of known onsets of symptoms per day as reported at different dates in the past. As testing confirms onset of symptoms in the past with varying delay, the epi curve not only grows at its tail, but over a wide time period with each new publication. Known onsets are reproduced from the RKI's daily situation reports and the publicly available RKI-database. Unknown onsets of symptoms, which account for 40% of total number of cases, are not considered. The estimated *total* epi curves from the RKI (imputation and Nowcasting), as reported on a past date, are not publicly available for the month of April, hence the focus on the numbers of *known* onsets here.

B. Differences between results based on RKI versus JHU data sources

At the beginning of the out break data were made available on a daily basis both by John Hopkins University (JHU) and the German Robert Koch Institute (RKI). Both sources initially provided only reported cases, with the JHU resources providing data faster and with a better interface for automated analyses. The RKI resources were updated only a few days later, as information always has to be transmitted from regional agencies to the RKI, whereas the JHU data for Germany are gathered from a few reputed online media (Berliner Morgenpost, Taggesspiegel and Zeit Online [8]). However the JHU resources have been partially criticised for lacking quality control (see issues section on the Github page [9]). We therefore compared the JHU data used in [1] to the official RKI count (Fig. 14) and have rerun the analysis using the RKI reported cases (the "Meldedatum", Fig. 15 and 16). The differences are minor.

V. IMPACT OF TESTING

Our modeling depends on reported case numbers, which in turn depend on testing. Throughout the COVID-19 spread, test availability, test requirements and known case numbers changed continuously over time, see Fig. 8. Such an inconsistent and fluctuating data-acquisition obviously introduces additional sources of uncertainty. While we decided to exclude the effects of testing in previous models, concerns about results derived from data that stem from inconsistent testing should be taken seriously. Thus, we analyze possible distortions in more detail. As we will demonstrate below, our major conclusions remain unchanged.

Please also note that at the time of writing of the initial manuscript, only very preliminary data and statistics on testing was available. Now, with better data, we come to the conclusion that reported case numbers, although they might derive from variable testing, are still useful to infer the actual disease dynamics.

In particular, evidence for the key characteristics of the first wave, i.e. strong exponential growth in new cases, change in transmission dynamics over a limited time period and slow exponential decline, can be derived from the available data, even if changes in testing are taken into account.

We start our analysis by considering two central quantities: i) the number of tests that are performed, say, on a given day or in a given week and ii) the fraction of the performed tests that are positive — a positive tests translates to a confirmed case.

Let us consider two simple limiting cases, in which only one of these quantities changes, whereas the other one remains constant. In the case that a constant number of tests is performed day-over-day and we observe a growing fraction of positive test results, this corresponds to an increase of the underlying case numbers. Conversely, if the number of tests is increased and we find a constant fraction of positive tests, this implies the same, an increase of underlying cases. The second case only holds with additional assumptions: i) the fraction of positive tests is larger than the prevalence and ii) tests are not performed randomly, both of which were met in Germany.

Fig. 9 A,B shows that in Germany in early March both, the number of tests as well as the fraction of positives increased simultaneously. This simultaneous increase indicates a significant growth in new case numbers.

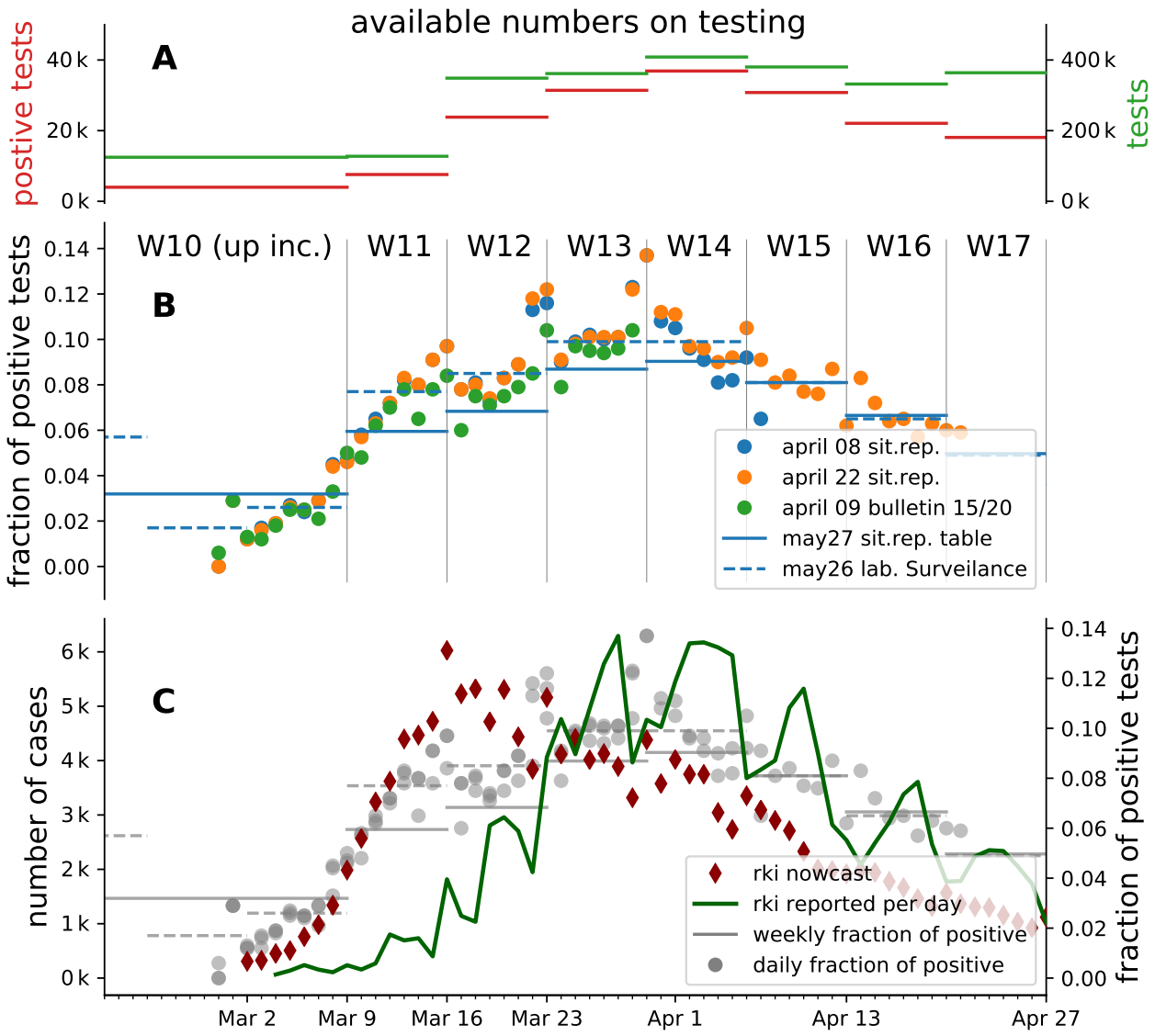

FIG. 9. The evolution of the fraction of positive tests in weeks 10 to 17. Weeks 10 to 12 strong exponential growth in in the number of new cases, which was not limited by the early testing capacity. **A** Comparison of number of positive test results with the number of tests performed for each week. Reproduced from Table 5 in [10]. Note: Numbers for week 10 and earlier are represented by a single data point. **B** Mid-term changes in the fraction of positive tests is more obvious in the daily data (points) than in the weekly (bars), especially in early March. Daily values are taken from situation reports [10–12] (full dataset) and the epi bulletin [13, 14] (ARS dataset). Weekly values, represented as horizontal lines, are taken from a situation report table and a weekly lab surveillance report (ARS dataset). Note: the latter represents a subset of all tests. Compared to the situation report, the ARS dataset lists weeks 8 to 10 individually. **C** Overlay of Panel 2 with the number of cases reported per day by the RKI and the estimated epi curve (imputation and Nowcasting, as described in [7]). The fraction of positive tests correlates with the number of reported cases from week 13 onward, as the total number of tests reaches a constant level.

A. Strong growth until week 12

Focusing on testing in weeks 10 & 11 in Fig. 9 A and B, we can clearly deduce a strong growth in daily new cases, as both the fraction of positives as well as the number of performed tests rise, matching the combination of the two scenarios described earlier. The rise in the fraction of positive tests is apparent in the daily values, especially as the daily number of tests can be taken as constant throughout the week, see Fig. 8 in [10]. For weeks 14 onward, the number of performed tests stays constant and thus, the fraction of positive tests correlates with the number of reported cases, exhibiting a decline in underlying case numbers. A similar direct comparison for weeks 10-12 is unfortunately not that simple, as the number of tests changed week-to-week in that time period.

For a better understanding of the following part of

the analysis, we recall an important fact on exponential growth: In each doubling period the same number of new infections occurs as in all preceding periods combined. As the number of tests approximately doubles every week until week 12 and the fraction of positive tests increases to week 13, the doubling period of new infections has to be shorter than 1 week. In a time frame of less than a week, more new infections occur than in the period since the onset of the outbreak. If we assume constant testing over the span of one week, a difference in the fraction of positive tests on each day during that week should be observable, and this is indeed what we see for testing in weeks 10 and 11 and to a lesser extent from start to end of week 12 (Fig. 9 B). A more in depth analysis of Fig. 9 is attached in Sec. V D. The important questions that remain are: When did the number of new infections peak? And when did it start to decline?

Deferring the first question to Sec. V B, we answer the second: From week 14 on, there is an approximately constant high level of testing, but a decline in the number of cases reported, and an accompanying day-to-day decrease in the fraction of positive test results. These observations are consistent with an exponential decline in the number of new infections, confirming that testing can properly measure the underlying epidemiological dynamics in this period.

Summing up the above analysis so far, we have indications that during the epidemic outbreak, a growth in case numbers was indeed present, as well as a decline.

Hypothetical Scenario: If we were to reject the above simple explanation that growing case numbers reflect growing numbers of infections, there is one alternative scenario to explain the observed trend, which we, however, deem highly implausible. As this scenario has frequently occurred in the public debate on the spread of COVID-19 in Germany, we discuss it briefly. The underlying assumption in this scenario is that the few tests that were performed during the initial outbreak until week 11 missed most of the actual cases, i.e. a large pool of infected persons would have existed unobserved. Then, at the same time at which the amount of tests was increased from weeks 11 to 12, coincidentally the effectiveness of the testing could have increased, so that the unobserved pool (of constant size!) is identified and, thus, apparent case numbers rise. Given the rigorous criteria (based on symptoms and risk of exposition) that were required from patients in order to qualify for one of the early tests, we deem this scenario of an unobserved and constant pool to be quite unlikely. Especially so because the fraction of positive tests stayed below 10% during the entire time.

B. Locating the peak position

In other words, we are interested in the peak position in the curve of onsets of symptoms, see again Fig. 9, C, red. The day of the peak is constrained by the initial simultaneous increase in tests and fraction of positive tests to occur no earlier than week 11, as the peak would indicate the end of the growth. In this section we're focusing on how it can be reliably identified from the stable period of testing: From week 12 and onward, the number of tests remained on an almost constant, high level $\sim$ 400k and changes in the *daily new cases reported* are directly reflected by the fraction of positive tests.

To understand this in more detail, we introduce the following important rule of thumb here: Tests of week i describe well what happened in week $i-1$.

The key is the connection between the date of symptoms onset (when symptoms first show), the testing (when the symptom onset is confirmed or an asymptomatic case is uncovered), and the reporting date (when a positive test-result is registered).

Any reported case must inherently be preceded by a test and according to the RKI, positive test results are reported within 24 hours to the responsible health department. The remaining task then is to reveal the connection between symptom onset and reporting date, i.e. the reporting delay for each individual case. The date of testing is taken as the day before reporting in the rest of the analysis, the testing delay is one day shorter than the reporting delay.

In Fig. 10 we detail the reporting delay by plotting distributions of *how many days after the symptom onset a case is reported*. For example, if each and every infected person would receive a test result (become a reported case) exactly three days after they showed symptoms, then the plotted distributions would have only one entry: a delta-peak at three days. However, we see that most

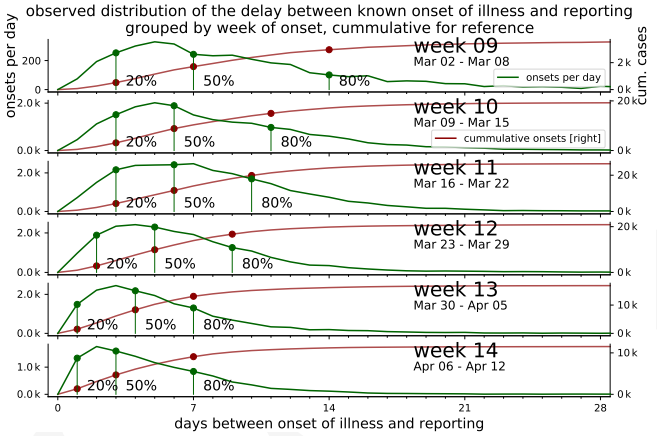

FIG. 10. The onsets of symptoms are confirmed by testing at later point in time, which accounts for most of the delay till all or the main fraction of known onset of symptoms (*IstErkrankungsbeginn* in RKI-database) are reported. From the RKI data, the number of cases per delay between onset of illness and reporting (i.e. *RefDatum* and *Meldedatum*) for cases with known onset of symptoms (*IstErkrankungsbeginn*) are counted for each week. The fraction of reported cases out of the total onsets up to a delay are highlighted for 20%, 50% and 80%. The cumulative number of cases reported up to each delay is displayed for reference.

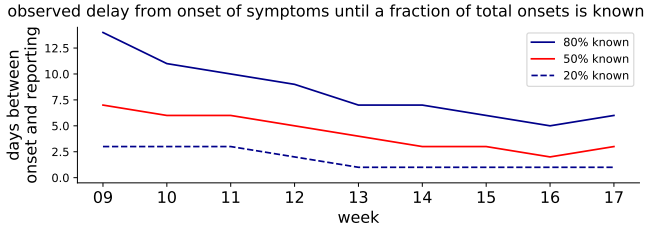

FIG. 11. Overview of the delay between onset of symptoms and the reporting of a fraction of total known onsets for a day changes with time. The 50% fraction represents the median reporting delay between onsets of symptoms and reporting. Derived from Fig. 10

reports arrive 1–7 days after symptom onset, where the details of the (lognormal) distribution depend on the week of onset of symptoms. Until and including week 12, the distributions have heavy tails. After week 12, the distributions have lighter tails. This provides some intuition of the distributions and the meaning of the heavy tails: *most* of the symptom onsets are reported within the first week but *some* will be reported much later, so that shape of the distribution still keeps changing. If the test level is low, *more* cases will be reported later and the tails of the distribution are heavier. This is latter effect is what we see for the onsets during the first weeks until 11; due to limited testing capacities, many cases are only reported weeks later — when more testing was available. To rephrase based on Fig. 11: Half of the onsets of symptoms in week 11 are reported within 5 days, 80% within 9 days. The crucial example here is: Half the onsets on Wednesday get tested until Sunday, the other half in the following weeks for every following day of the week the fraction of test performed in the next week rises. Without explicitly working out the details, it's fair to declare the initial rule of thumb valid. A more thorough analysis based on actual per case testing-delays instead of reporting delay distributions is conducted in Sec. V C.

Let's turn back to Fig. 11 A. The onsets in week 11, the estimated position of the peak, should be robustly measured by the testing in week 12, with a high number of total tests. From Fig. 11 C, we can see, that the number of onsets of illness peak at the end of week 11 or the beginning of week 12. This time point doesn't suffer from lower testing numbers in week 11.

C. Decomposing the epi curve into weeks of testing

Having established the delay between symptom onset and reporting, we can decompose the epi curve and identify parts of the curve that stem from certain weeks of testing. Fortunately the publicly available RKI database contains both onsets and reporting for individual cases for 60% of the total cases and thus also the date of testing, which in general is one day earlier than the report. In more detail, for all the cases within a chosen testing period, we also know the respective date of onset of symptoms, for complete datasets. Borrowing from [7], the remaining 40% of test dates can be imputed from the known onsets dataset. In Fig. 12 A,B we apply this method to collect all the symptom onsets that were found by testing in weeks 12 and 13. Through this allocation of "which part of the curve stems from which tests", we can thoroughly justify the connection that we made above, when we said that growth in weeks 11 and 12 stems from the tests in week 12 and 13. As we see, the peak on March 16 stems almost completely from tests of week 12 and 13; these weeks already featured the high level of tests performed. Based on the decomposition, we can conclude that in week 11, every day could have been identified as the peak based on testing in weeks 12 and 13.

We can extend this method in an attempt to reduce the influence of changing number of tests per week on the estimation of the change in the number of onsets of symptoms from one week to the next. We compare the number of onsets in different weeks, that were confirmed by one week of testing. Think: *distribution of onsets per week seen by the testing in one single week*. Some cases with onset of symptoms on Monday will receive their positive result within the same week as the symptom onset itself, others get tested further away from their onset of symptoms. As viewed from one single week of testing, we distinguish 4 categories: onsets 3 weeks, 2 weeks and 1 week earlier than the test and onset in the same week as testing. The number of onsets in each of the 4 categories compared with the total number of onsets confirmed in the week of testing, the fraction per category, is characteristic for the epidemiological dynamic in the time span of those four weeks. This method is more robust to changes in the number of tests week-over-week, than the other methods outlined so far. In Fig. 13 three different scenarios are considered and their effect on the fraction of cases in each week-category is worked out. All three scenarios show distinctive combination of fractions per week-category. Comparing the artificial result with Fig. 12 C, we find that in week 11 most of the tests (52%) found symptom onsets within the same week. This indicates weeks 10 and 9 had significantly less new onsets of illness. This is consistent with the exponential growth uncovered in sec. V A. In the extreme case that no tests were performed in week 10 and we were to observer that the number of onsets in week 10 were comparable or higher than in week 11, the backlog from week 10 would lead to higher fraction of *1-week-earlier* onsets than same week onsets, for testing in week 11. As the fraction of 1-week-earlier onsets is lower than same-week for testing in week 11, we can see that the assumption, no tests and highter number of cases in week 10, cannot be valid. Reaffirming the observation of growth from week 10 to week 11. Testing in week 12 shows a significant peak for onsets 1 week earlier. That indicates the number of new onsets is comparable in week 11 and week 12 (see artificial result, Fig. 12). Note, that a lower total number of tests in week 11 amplifies this observation. Weeks 13 onward

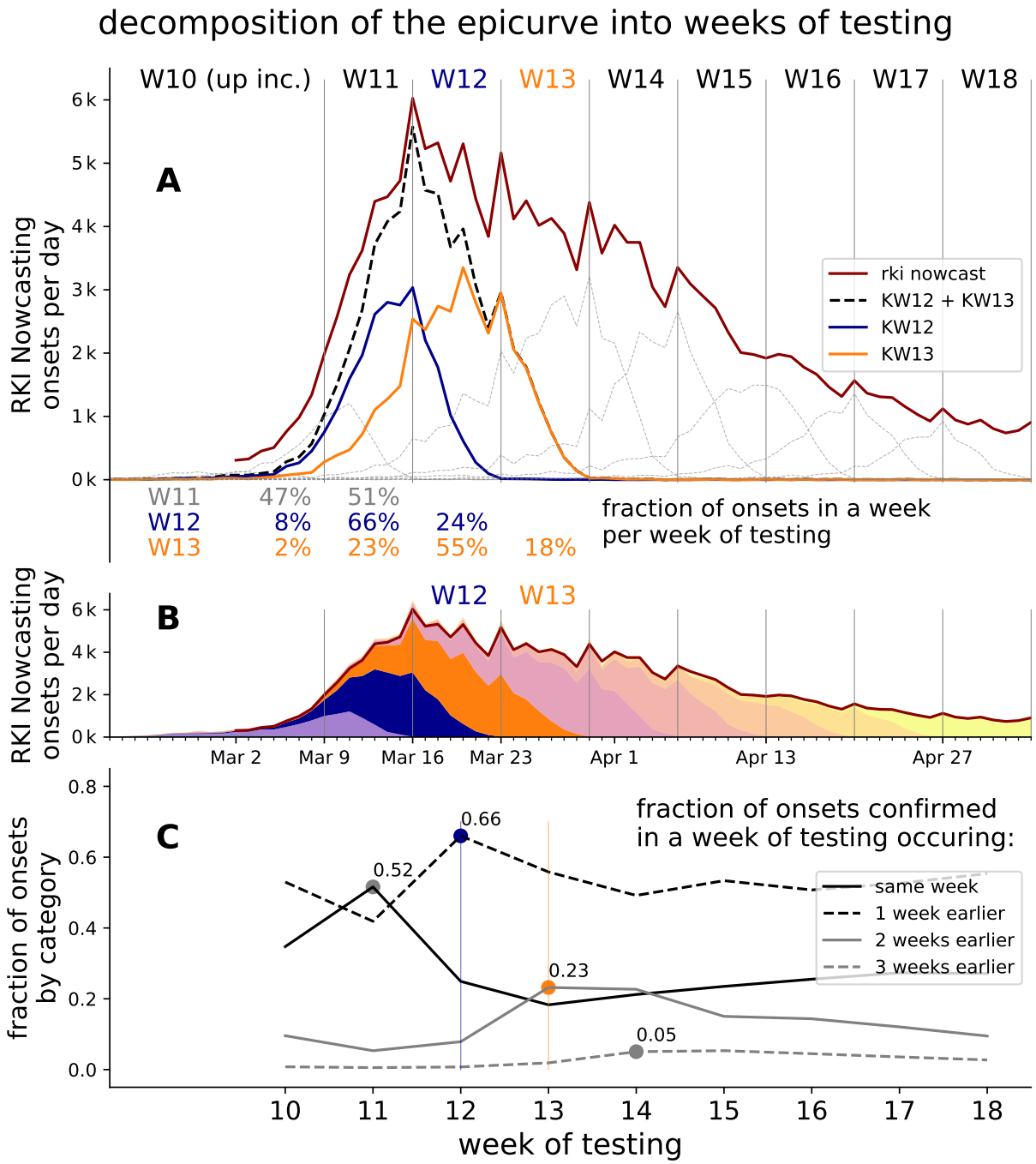

FIG. 12. Testing in one week confirms onsets of symptoms that occur up to 4 weeks earlier. The extend of this effect is analyzed based on the RKI database through decomposition by allocation of onsets of symptoms to weeks of testing. It is assumed that the delay between the time of testing and *Meldedatum* is 1 day. Tue-Mon *Meldedatum* is taken as a proxy for Mon-Sun testing. **A** Onsets of symptoms per day curves allocated to weeks of testing, weeks 12 and 13 are highlighted. Most known onsets around the peak of the epi curve in week 11 are confirmed by the testing in weeks 12 and 13. **B** stacked decomposition of the epi curve into weeks of testing. **C** To reveal crucial information about week-to-week change in the number of total onsets based on one week of testing, the shape of the distributions of onsets of symptoms confirmed by that week of testing is characterized. The fraction of onsets in the same week and each preceding week out of the total onsets confirmed by the week of testing is calculated. This indicates, the portion of a week's positive tests confirming onsets in the same week or in preceding weeks (max. 3 weeks earlier). The evolution of these 4 values is plotted by the week of testing. The peak of the epi curve can be tracked through testing results of weeks 11 to 14 as a maximum in the same-week/n-weeks earlier fraction of onsets confirmed in those respective weeks: 52% of all cases confirmed through testing in week 11 had onset of symptoms in the same week. Even more notable: 66% of positive tests in week 12 are linked to onsets 1 week earlier: in week 11. For comparison, see Fig. 12

show distributions which indicate decline in onsets week over week, their 2 weeks earlier fractions are larger, while their fraction of same-week onsets is smaller than 30%.

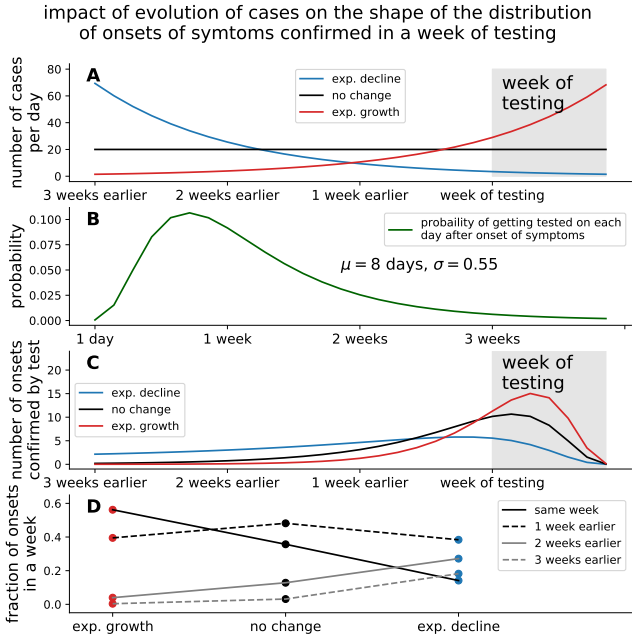

FIG. 13. Changes in the number of onsets of symptoms from one week to the next can be estimated from the distribution of onsets of symptoms confirmed by testing in the latter week, if we group those onsets by week of onset. **A** Three different scenarios for the evolution of the number of cases are considered, whereby the number of onsets of symptoms per day is plotted. **B** Each case from A has a probability to be tested on every day. Half of the cases get tested within 8 days. The distribution is derived from observed data. **C** Number of onsets confirmed by the week of testing for each day of onset of symptoms. As a result of A and B. The shape of the distribution is characteristic for the change in cases and can be compared with Fig. 12, B. **D** As a last step the onsets confirmed by the testing in the highlighted week are summed up by week of onset and the group's respective fraction of the total number of positive tests in the highlighted week is computed. If no change in the number of cases occurs, more onsets in the week preceding testing are confirmed (45% of total) than from the same week as testing (35% of total). In case the number of cases rises, onsets from the same week as testing constitute the majority of onsets confirmed by tests in the week. If the number of cases declines, old onsets (older than 1 week) take over a significant fraction of total onsets tested in the week.

In Summary: Even though the number of total tests performed changed until week 12, the available data indicates **strong** exponential growth in new onsets of symptoms into week 11, constraining the peak in new onsets of illness to no earlier than March 9. The declining phase of the wave is well documented. The exponential decline in cases from week 13 onward is measured with consistent high level in the number of tests. As testing in one week is shown to uncover onsets of symptoms in the 3 preceding weeks, the alleged period of the peak in new onsets of symptoms in week 11 is covered by robust testing from weeks 12 and 13. Based on testing in weeks 12-13, the peak can be identified at the end of week 11 or beginning of week 12.

D. Available data on testing

The epi bulletin [15] outlines the different networks that the RKI uses to source information on testing: *Voxco*, *RespVir*, the antibiotics-resistance-surveillance (ARS) [14] and lab-accociation queries. These sources are compiled into weekly data-sets with total number of tests and positive tests, which are published in the daily situation report once a week.

Data from the ARS contains daily number on testing and a separate weekly report is published on the RKI website. The ARS dataset covers 25-30% of the total number of tests reported by the RKI, as only 62 of 180+ labs participate. The ARS data-set shows a mean delay between sampling and testing between 1 and 1.2 days except for weeks 12 to 15, where the delay is 1.5 days, peaking in week 13 at 1.8 days.

An overview of all publicly available data on testing for march 2020 is presented in Fig 9. The following observations along with additional comments are based on this presentation:

- From week 8 to week 12 the number of tests rises week to week by a factor greater than 2. 120k is a combined number for weeks up to 10. Individual numbers of tests for those weeks has to be estimated with help from the ARS-subset (Fig. 9 B *may26 lab. Surveilance*). Assuming ARS is representative the number of test performed in week 10 should be around 60k, 30k in week 9 and 30k in all weeks up to and including 8, extending the exponential pattern.

- The number of tests remains on a high level from week 12 on. In the range of 340-430k.

- The number of positive test rises faster than the total number of tests until week 14.

- The fraction of positive tests per week peaks around 10%, relatively low compared with neighbouring countries.

- The fraction of positive tests per day varies with time from 2% around March 1 to around 10% in weeks 13 and 14, peaking at 14% at the end of March. Afterwards declining to less than 2% in week 20 (not shown in figure). The day-to-day rise in week 10 and 11 is more pronounced than the weekly average would suggest.

- The increase in the fraction of positive tests does not correlate to the rise in number of reported cases

until week 13, but correlates with the decline in reported cases from week 13 on, which is expected as the total number of tests fluctuates around 380k tests per week on a high level. The correlation with the epi-curve is coincidental.

- The ARS data shows a steady day to day increase in positive fraction of test in weeks 10 and 11. Weekends show a higher fraction, while the total number of tests is lower (daily total number not shown in the figure). Deviating from the rise in the positive fraction, weeks up to 8 have a 3 times higher fraction of positive results than week 9.

- The maximum test-capacity per week as reported by the labs increased to 1M in week 19, showing strong growth till week 14. A week to week doubling in test capacity continues for two more weeks compared to growth in number of tests performed (not shown).

Additional information relevant to the discussion can be found in the publications cited earlier. For the total data-set, the fraction of positive tests varies from 1.5 to 7.2% for different states. Not a single day of testing for individual states exceeded 20% positive results.

VI. SUMMARY & CONCLUSIONS

In these technical notes, we have comprehensively addressed questions and comments regarding our recent publication [1]. First, we compared direct, model-free estimates of the reproduction number to the ones obtained from dynamical modeling. To this end, we established synthetic ground-truth data based on an SIR model and subsequently inferred the reproduction number based on various complementary approaches that are in practical use. We reveal how sudden changes in the spreading rate, as expected from the broad implementation of non-pharmaceutical interventions, can lead to counterintuitive transient drops in new reported cases. Most importantly, we find that only modeling of spreading dynamics can correctly capture effects of sudden changes in the spreading rate.

Second, we provided extensive background on our modeling rationale which combines differential-equation based modeling of dynamics with Bayesian parameter inference and formal model comparison. Within the Bayesian framework, we argue that based on prior knowledge, the most plausible models explaining the data can be systematically identified and also updated as new information becomes available. We also discuss why we do not think that models based on herd immunity are plausible given our present knowledge.

Third, we analyzed additional data on the COVID-19 spread in Germany, which has become available since the completion of the analysis presented in [1]. Most importantly, we include data sets from the German Robert Koch Institute based on the reporting date as well as based on the onset of symptoms (epi curve). We analyzed the data in the framework of SIR and SEIR models, and we also tested a broad range of varying prior assumptions. We find our results to be robust across these varying modeling assumptions and data sets, and to support the conclusions drawn in [1]. In turn, this leads us to conclude that under the conditions comparable to those in Germany, models based on reporting date are a viable alternative for analyzing the early stages of a disease outbreak, before the epi curve becomes available — as long as the reporting delay is properly modeled.

Finally, we address the issue of changes in the testing capacities and procedures over the course of our analysis. Most importantly, we find that while data from the initial onset of the pandemic is presumably affected by a rise in test capacities, the crucial part of our analysis is based on a regime of comparably stable testing. In particular, we find that the inference of the second and third change point is unaffected by testing.

Overall, the analysis here evaluates the robustness of our previously reported results with respect to statistical and dynamical modeling assumptions as well as complementary data sources and provides additional support for the central conclusions of our publication [1].

[1] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann. Inferring change points in the spread of covid-19 reveals the effectiveness of interventions. *Science*, 2020.

[2] Theodore Edward Harris. *The Theory of Branching Processes*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin Heidelberg, 1963.

[3] https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/R-Wert-Erlaeuterung.html.

[4] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2nd edition edition, September 2009.

[5] Nihat Ay and Daniel Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(01):17–41, February 2008.

[6] Michael Höhle and Matthias an der Heiden. Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002, 2014.

[7] M. an der Heiden and O. Hamouda. Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland – Nowcasting. *Epid. Bull.*, 2020.

[8] Tagesschau.de. Exklusiv: Woher die Johns-Hopkins-Zahlen zu Corona stammen, https://www.tagesschau.de/inland/johns-hopkins-uni-corona-zahlen-101.html.

[9] CSSEGISandData. COVID-19, June 2020. original-date: 2020-02-04T22:03:53Z.

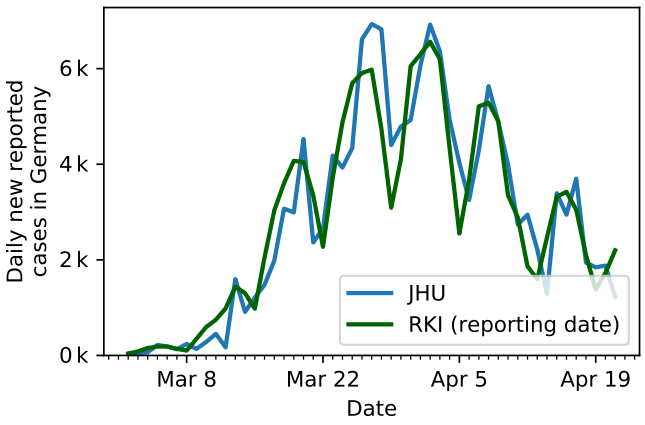

FIG. 14. Comparison of the German case numbers as published by the John Hopkins University (JHU) used in our previous publication [1], to the case number of the Robert Koch Institute (RKI). The difference is limited.

[10] Täglicher Lagebericht des RKI zur Coronavirus-Krankheit-2019 2020-05-27, 2020.
[11] Täglicher Lagebericht des RKI zur Coronavirus-Krankheit-2019 2020-04-22, 2020.
[12] Täglicher Lagebericht des RKI zur Coronavirus-Krankheit-2019 2020-05-22, 2020.
[13] A. Hoffmann, I. Noll, N. Willrich, A. Reuss, M. Feig, M.J. Schneider, T. Eckmanns, O. Hamouda, and M. Abu Sin. Laborbasierte Surveillance SARS-CoV-2. *Epid. Bull.*, 2020.
[14] SARS-CoV2-Surveillance - Wochenbericht vom 26.05.2020, 2020.
[15] J. Seifried and O. Hamouda. Erfassung der SARS-CoV-2 Testzahlen in Deutschland. *Epid. Bull.*, 2020.

VII. SUPPLEMENTARY INFORMATION: FIGURES

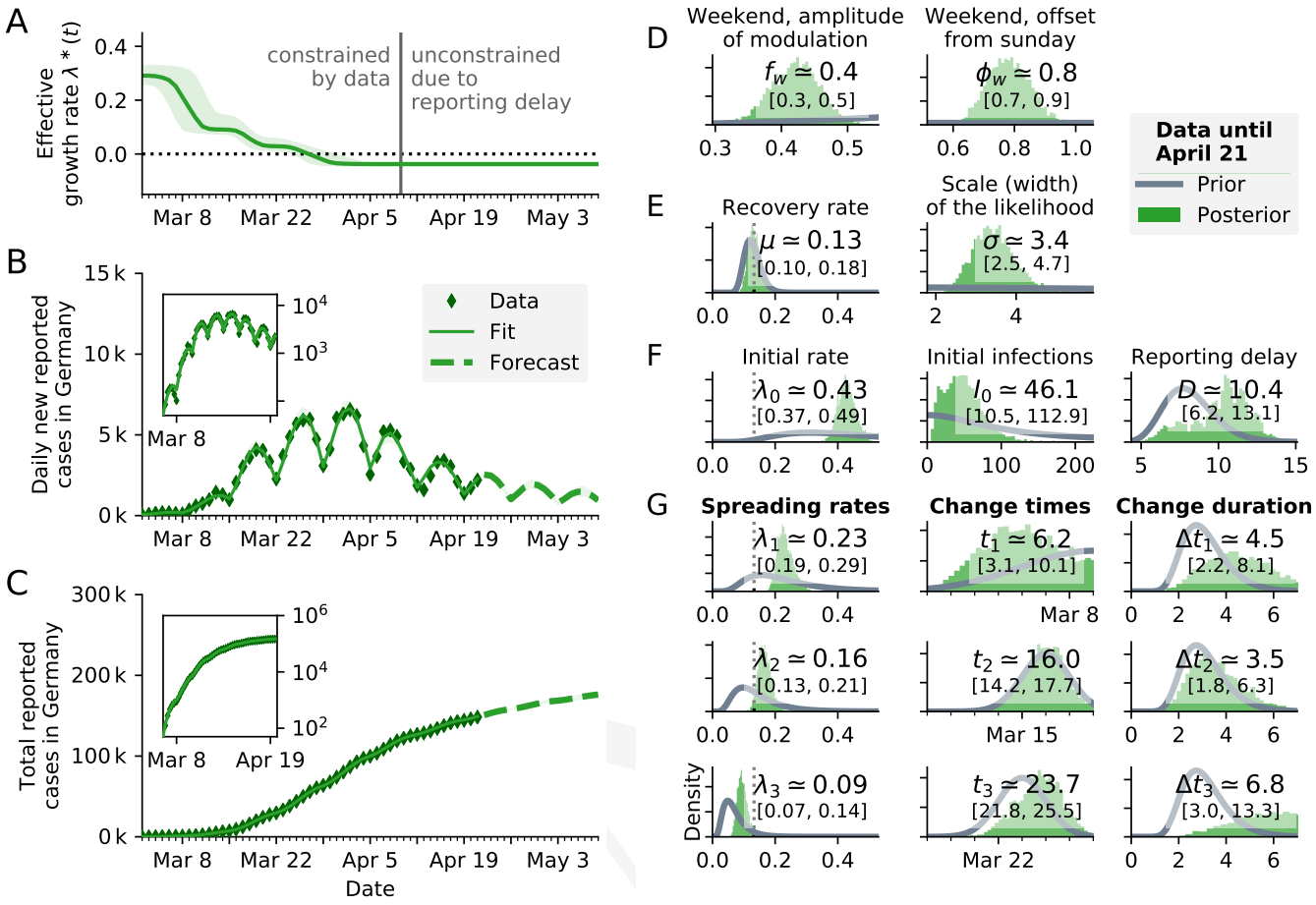

FIG. 15. **SIR model** (see Fig. 3 of [1]) using the **reporting date (Meldedatum) of the RKI data** for inference. **A** Time-dependent model estimate of the effective spreading rate $\lambda^*(t)$. **B** Comparison of daily new reported cases and the model (green solid line for median fit with 95% credible intervals, dashed line for median forecast with 95% CI); **inset** same data in log-lin scale. **C:** Comparison of total reported cases and the model (same representation as in B). **D–G** Priors (gray lines) and posteriors (green histograms) of all model parameters; inset values indicate the median and 95% credible intervals of the posteriors.

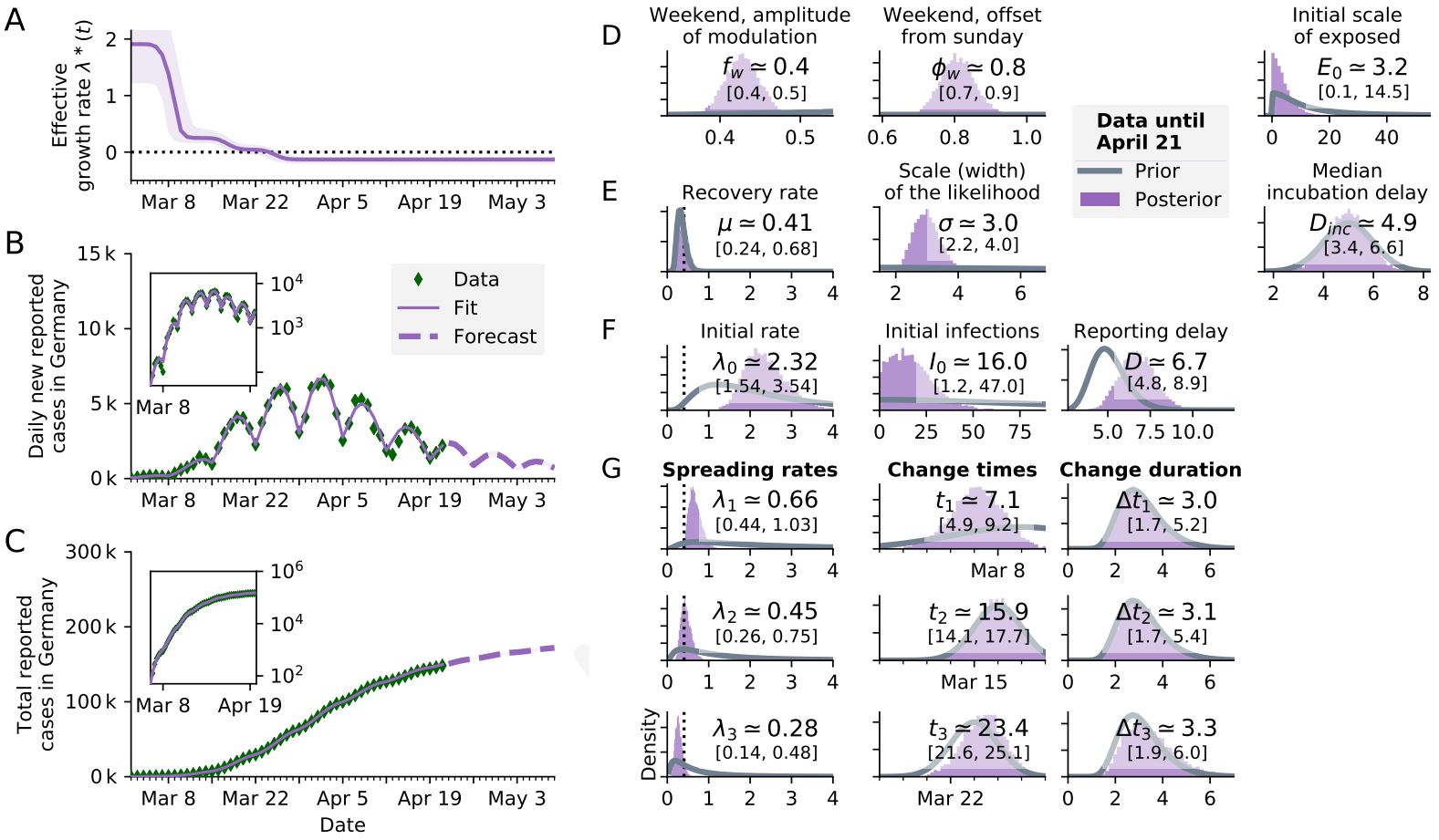

FIG. 16. **SEIR-like model** (see Fig. S3 in Supplementary Information of [1]) using the **reporting date (Meldedatum) of the RKI data** for inference. **A** Time-dependent model estimate of the effective spreading rate $\lambda^*(t)$. **B** Comparison of daily new reported cases and the model (purple solid line for median fit with 95% credible intervals, dashed line for median forecast with 95% CI); **inset** same data in log-lin scale. **C** Comparison of total reported cases and the model (same representation as in B). **D**–**G** Priors (gray lines) and posteriors (purple histograms) of all model parameters; inset values indicate the median and 95% credible intervals of the posteriors.

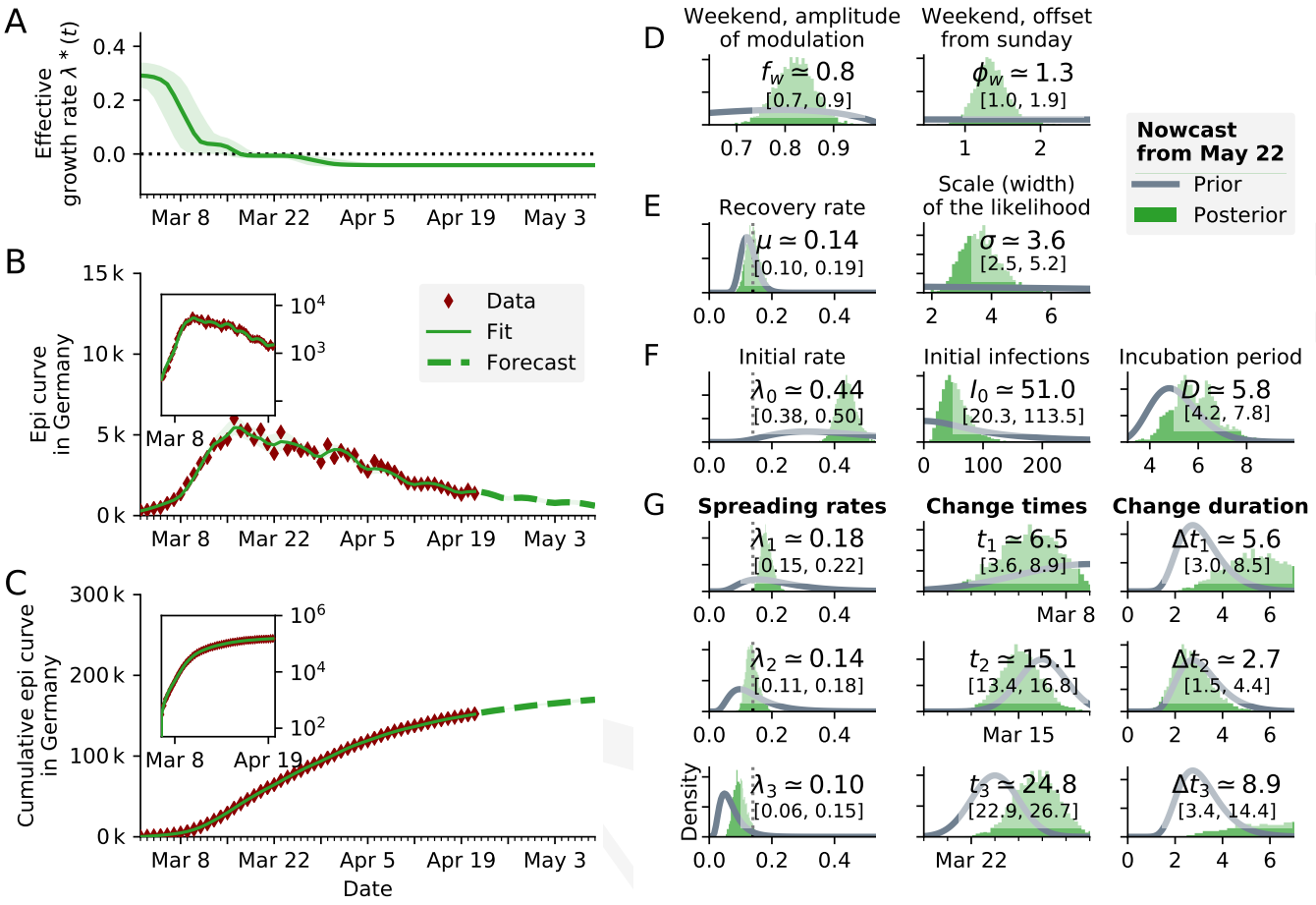

FIG. 17. **SIR model using the onset of symptoms** (nowcast from May 22) of the RKI data for inference. The median of the lognormal prior of the delay between infection and onset of symptoms has been set to 5 days (right-most panel F). **A** Time-dependent model estimate of the effective spreading rate $\lambda^*(t)$. **B** Comparison of daily new reported cases and the model (green solid line for median fit with 95% credible intervals, dashed line for median forecast with 95% CI); **inset:** same data in log-lin scale. **C** Comparison of total reported cases and the model (same representation as in B). **D**–**G** Priors (gray lines) and posteriors (green histograms) of all model parameters; inset values indicate the median and 95% credible intervals of the posteriors.

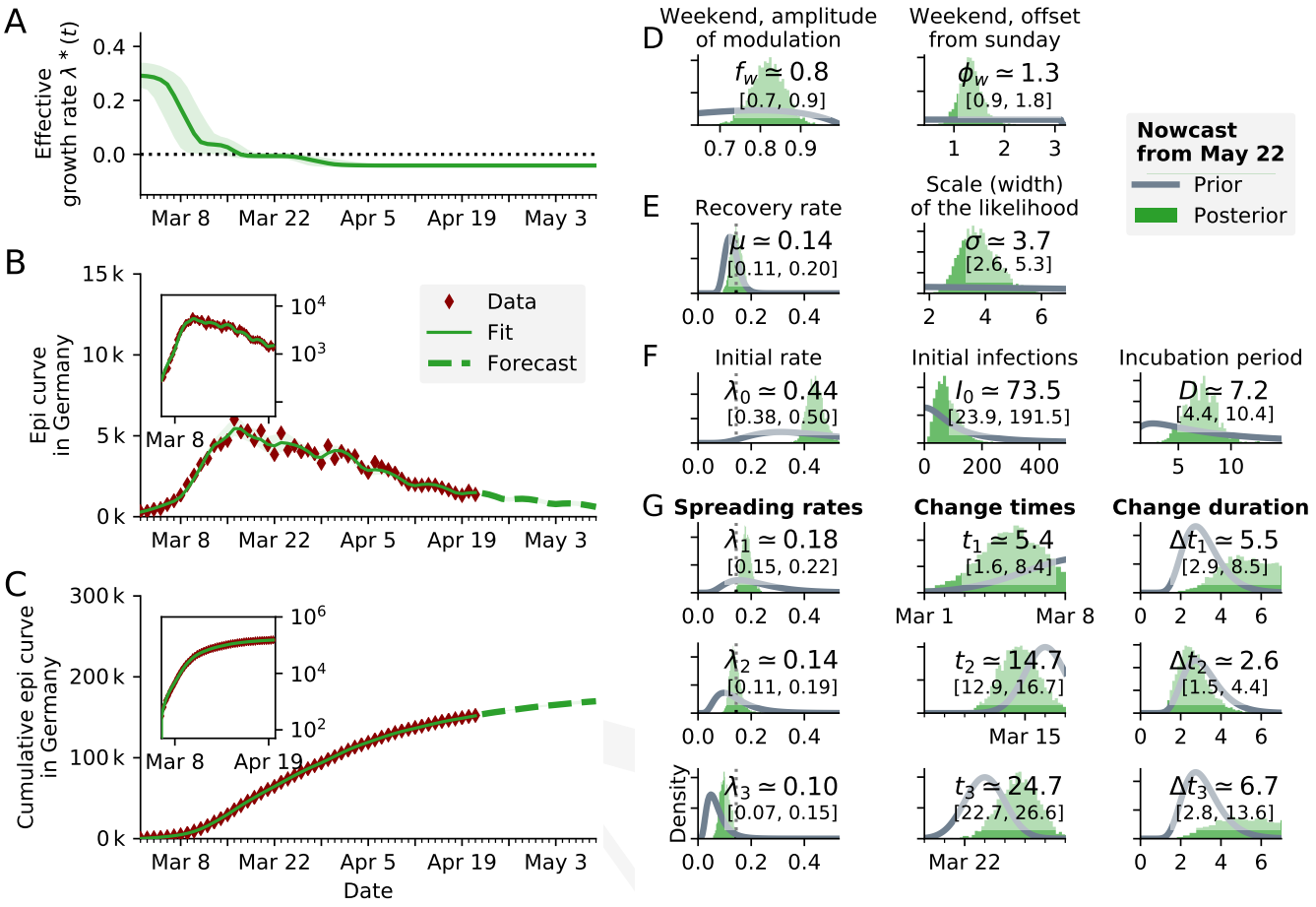

FIG. 18. **SIR model using the onset of symptoms** (nowcast from May 22) of the RKI data for inference. **The median of the lognormal prior of the delay between infection and onset of symptoms has been set to a relatively uninformative prior** (right-most panel F). The posterior of the delay has as median 7.2 days, which is close to the expected incubation period of 5 days. **A** Time-dependent model estimate of the effective spreading rate $\lambda^*(t)$. **B** Comparison of daily new reported cases and the model (green solid line for median fit with 95% credible intervals, dashed line for median forecast with 95% CI); **inset** same data in log-lin scale. **C** Comparison of total reported cases and the model (same representation as in B). **D–G** Priors (gray lines) and posteriors (green histograms) of all model parameters; inset values indicate the median and 95% credible intervals of the posteriors.

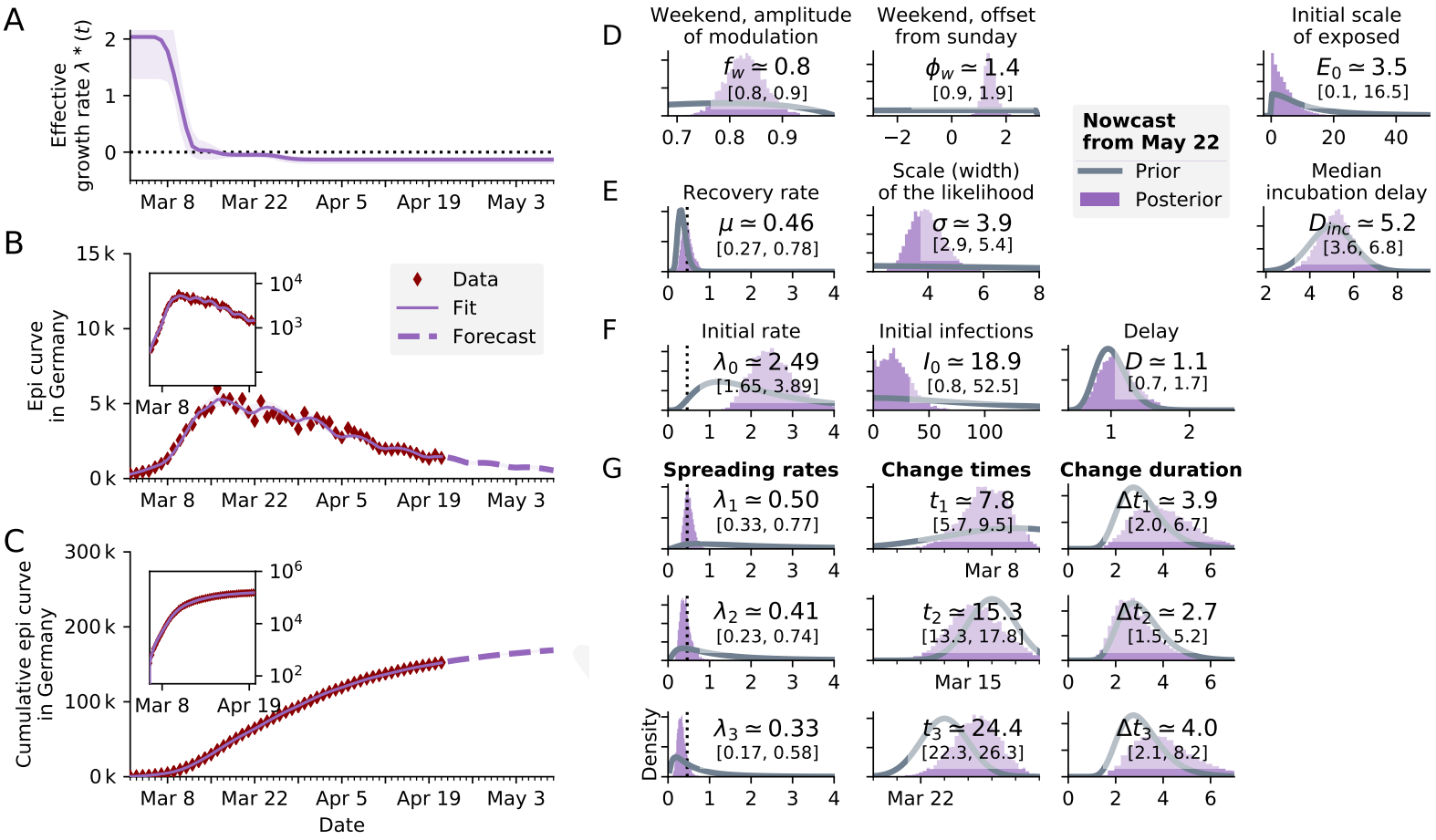

FIG. 19. **SEIR-like model using the onset of symptoms** (nowcast from May 22) of the RKI data for inference. The median of the lognormal prior of the delay between infectious and onset of symptoms has been set to 1 day (right-most panel F). **A** Time-dependent model estimate of the effective spreading rate $\lambda^*(t)$. **B** Comparison of daily new reported cases and the model (purple solid line for median fit with 95% credible intervals, dashed line for median forecast with 95% CI); **inset** same data in log-lin scale. **C** Comparison of total reported cases and the model (same representation as in B). **D**–**G** Priors (gray lines) and posteriors (purple histograms) of all model parameters; inset values indicate the median and 95% credible intervals of the posteriors.

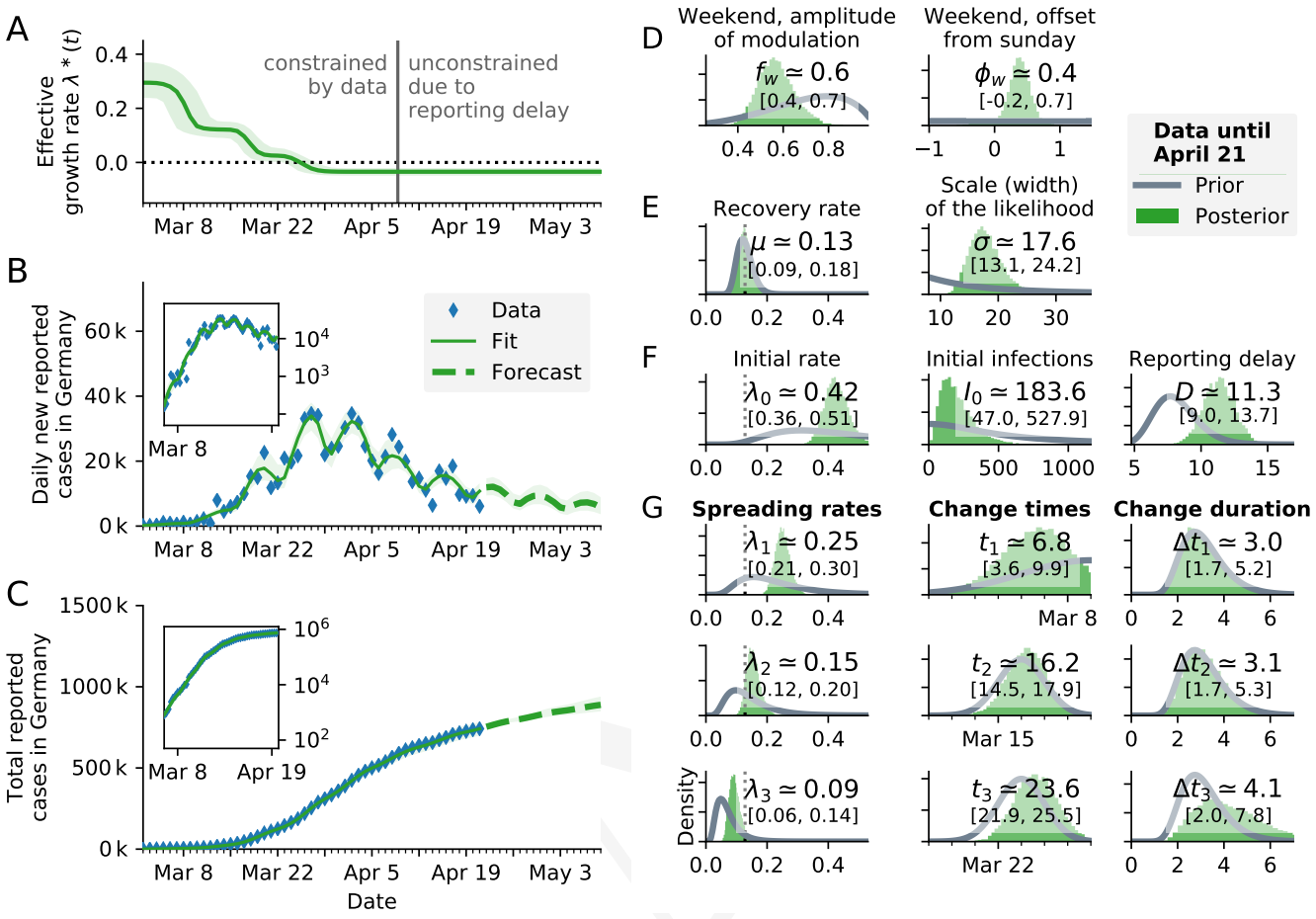

FIG. 20. SIR model **with reported case number multiplied by 5, to account for an eventual factor five of unknown cases**. Results are nearly identical to original non-multiplied plot (Fig 3. in [1]), showing that a constant underreporting has a negligible effect. The median inferred spreading rates λ are about 0.01 larger. **A** Time-dependent model estimate of the effective spreading rate $\lambda^*(t)$. **B** Comparison of daily new reported cases and the model (green solid line for median fit with 95% credible intervals, dashed line for median forecast with 95% CI); **inset** same data in log-lin scale. **C** Comparison of total reported cases and the model (same representation as in B). **D**–**G** Priors (gray lines) and posteriors (green histograms) of all model parameters; inset values indicate the median and 95% credible intervals of the posteriors.

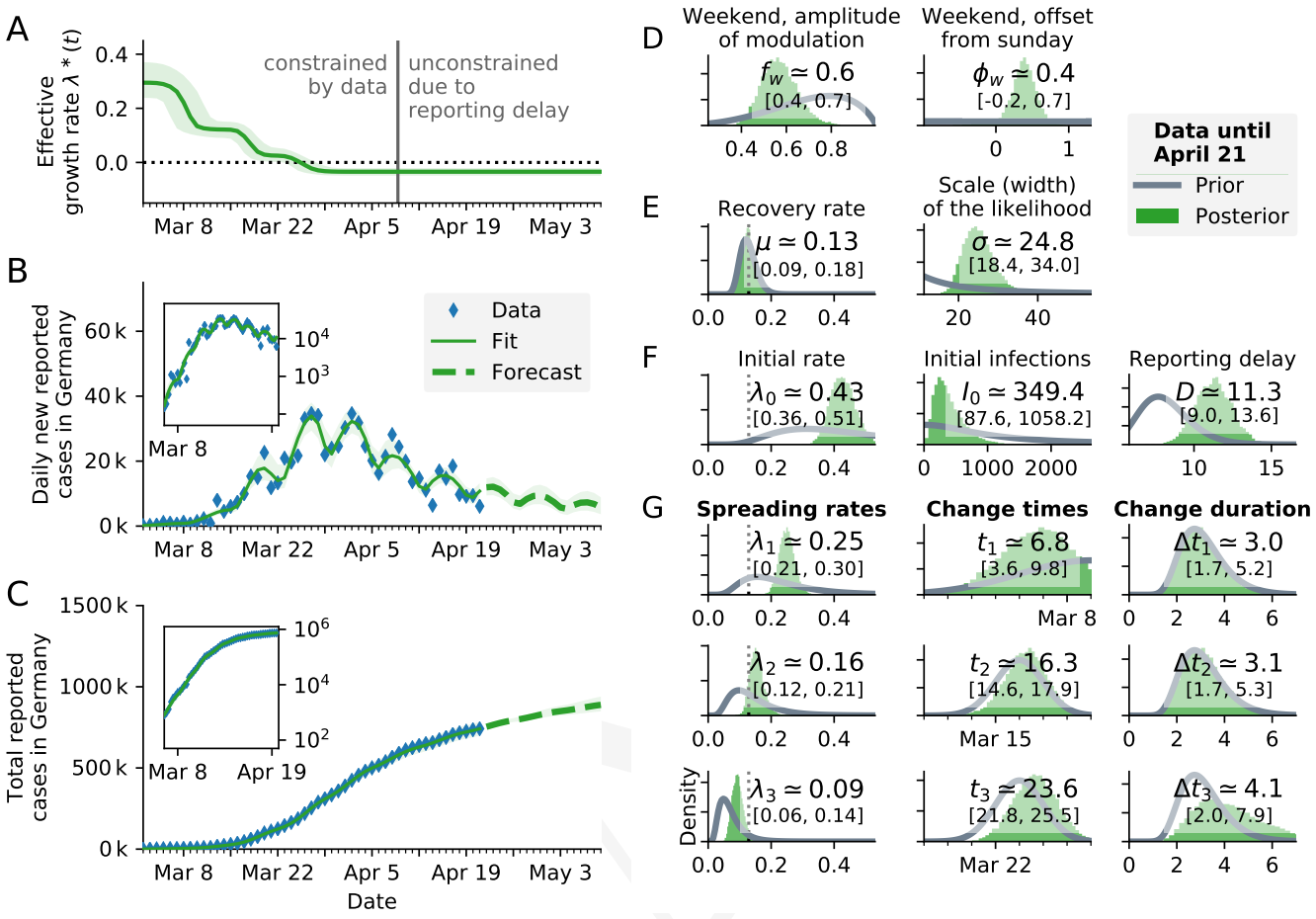

FIG. 21. SIR model **with reported case number multiplied by 10, to account for an eventual factor 10 of unknown cases.** Results are nearly identical to original non-multiplied plot (Fig 3. in [1]), showing that a constant under-reporting has a negligible effect, similar to Fig. 20. The median inferred spreading rates λ are 0.01-0.02 larger. **A** Time-dependent model estimate of the effective spreading rate $\lambda^*(t)$. **B** Comparison of daily new reported cases and the model (green solid line for median fit with 95% credible intervals, dashed line for median forecast with 95% CI); **inset** same data in log-lin scale. **C** Comparison of total reported cases and the model (same representation as in B). **D–G** Priors (gray lines) and posteriors (green histograms) of all model parameters; inset values indicate the median and 95% credible intervals of the posteriors.