medRxiv preprint doi: https://doi.org/10.1101/2020.05.12.20099366; this version posted May 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

A pitfall in estimating the effective reproductive number R_t for COVID-19

Markus Petermann[a] and Daniel Wyler[b]

[a]Zentralstrasse 47, CH-8003 Zürich, Switzerland
[b]Institut für Theoretische Physik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

May 12, 2020

Abstract

The effective reproductive number R_t of COVID-19 is determined indirectly from data that are only incompletely known. Approaches based on reconstructing these data by sampling time lags from suitable distributions introduce noise effects that can result in distorted estimates of R_t. This, in turn, may lead to misleading interpretations of the efficacy of the various measures taken to limit COVID-19 transmission. We discuss in some detail a study used for real time monitoring of the reproductive number in Switzerland; see https://bsse.ethz.ch/cevo/research/sars-cov-2/real-time-monitoring-in-switzerland.html.

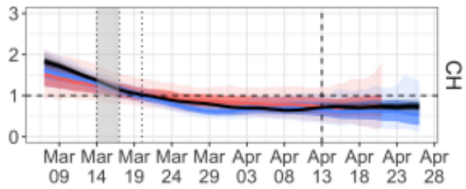

Figure 1: Plot from the real time monitoring of R_t in Switzerland on May 8. The shaded region is the onset of strong public health measures.

We argue that the method used to derive the above curve is systematically flawed and leads to an underestimation of the efficacy of the lockdown. The method adopted by the Robert Koch Institute suffers from similar deficiencies, their impact is however smaller.

Introduction

The daily varying effective reproductive number R_t is often used to monitor the spread of epidemic diseases such as COVID-19. It measures the expected

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.12.20099366; this version posted May 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

number of secondary infections on day t due to a single infected individual and is given by

$$R_t = \frac{I_t}{\sum_{n>0} I_{t-n} w_n}$$

where I_t is the number of new infections on day t (and equally I_{t-n} the number of infections on day $t-n$ and so on), and the w_n are the infection intensities, i.e. w_n is the average percentage of secondary infections caused on a given day by a person who was infected n days earlier. While the infection intensities can be fitted to available data, the I_s are not observable directly, unless representative proportions of the population were tested on a daily basis. Therefore, they need to be inferred indirectly from some other data.

There exist different schemes to reconstruct the I_s from the data, like the classical statistical inference methods. In the present article we concentrate on schemes that are based on the idea of a "mechanical" reconstruction of the data by sampling time lags of the observed data from suitable distributions. In the context of the current COVID-19 pandemic, such schemes have been implemented in different ways by various groups in different countries.

In the following section, we present these schemes and show how they systematically introduce noise into the true data. In the remaining sections we examine the impact of the noise on the reproductive numbers calculated from these data.

The reconstruction scheme and its smoothing effect

We exemplify the scheme and how it introduces noise into the data by the version implemented by the Computational Evolution group at ETH Basel [1]. For the sake of clarity, we limit our exposition of the scheme to the observables C_s, the number of confirmed cases on day s. The other observables used analogously by the group are hospitalizations and deaths. Moreover, as our focus is on the days around March 17, when the lockdown started, and as close to two month have passed since then, we restrict our exposition to those days of infection where all infections can be assumed to be confirmed by the actual date of the monitoring. According to the parameters used by the group, 95% of the cases are confirmed within 20 days after infection. For their method of extending the reconstruction to later days we refer again to their article.

Now, let X denote the incubation time of a randomly drawn case, i.e. the time between infection and symptom onset, and analogously Y the time between symptom onset and confirmation. The distributions of X and Y result from fitting to available data; see [1] and references therein. Then for every confirmed case a one samples independently a x'_a from the distribution of X and a y'_a from the distribution of Y.

The *reconstructed* infection day i'_a of this case a is then simply the day when the virus infection was confirmed minus $(x'_a + y'_a)$. Counting the number of cases that fall now on day s gives the reconstructed I_s, that we denote by I'_s. The reproductive numbers calculated from these I'_s are denoted by R'_s.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.12.20099366; this version posted May 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

That this scheme introduces noise into the true data is seen as follows. We denote the true infection day by i_a, the true incubation period by x_a and the true time between symptom onset and confirmation by y_a. Then we have $i'_a = i_a + x_a + y_a - x'_a - y'_a$. As the sampled x'_a and y'_a are independent of the true x_a and y_a (because we don't know these; we just know that they are approximately distributed like X and Y, respectively), the reconstructed i'_a equals the true "signal" i_a plus some "noise" $d'_a = x_a + y_a - x'_a - y'_a$.

As we will see in the now following examples, this results in a smoothing of the infection number statistics, which in turn, under certain circumstances, has a significant impact on the reproductive numbers calculated from it.

An illustrative example

The following example illustrates the effect of this scheme on the reconstructed reproductive number. Assume that $I_1 = 128$ and $R_s = 2$ for $s \leq 6$ and $R_s = 0.8$ for $s \geq 7$, and that infectiousness is limited to the day after infection, i.e. $w_1 = 1$. This yields the 'true' infection numbers and reproductive numbers which are illustrated by the read curves in Fig.2. For the *reconstructed* data we take X and Y both to be gaussian with mean 5 and standard deviation 1. Thus, the "noise" is also gaussian with mean 0 and standard deviation 2. With this "noise", the scheme results in the average in the corresponding blue curves.

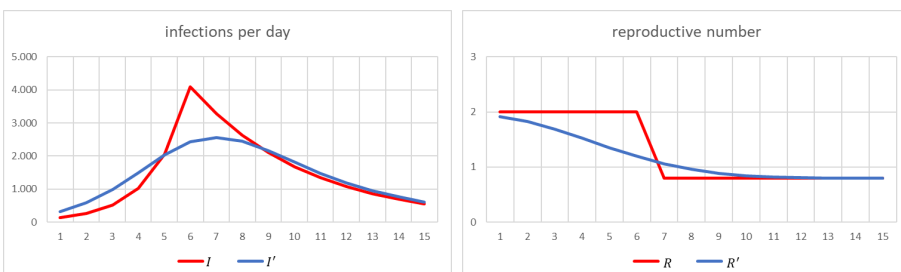

Figure 2: 'True' (red) and reconstructed (blue) infection and reproductive numbers

Of course, the blue curve R' is prone to lead to wrong decisions: If a lockdown caused the sharp decline in R from day 6 to day 7, then the blue curve may suggest that its impact was much less important and that most of the reduction was achieved already before the lockdown. It might even lead to the conclusion that the lockdown was not needed at all and that softer measures in force already before day 6 have had a sufficient effect, where in reality they had no effect at all. This in turn could lead to the conclusion, that the pandemic can be kept under control by adhering to soft measures only.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.12.20099366; this version posted May 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

The reproductive number in Switzerland around March 17

A calculation based on distributions for the time lags X and Y and the infection intensity as described in [1] yields the following result:

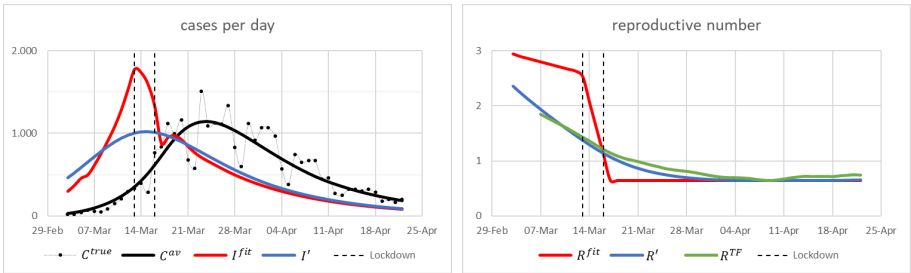

Figure 3: Cases and reproductive numbers in Switzerland around March 17. See text below for the definition of the curves.

Unlike in the illustrative example above, we cannot start from 'true' numbers of new infections. We instead choose the numbers of new infections, denoted by I^{fit} (red curve), in such a way that the resulting expected numbers of confirmed cases C^{av} (solid black curve) fit well the black dots C^{true} which show the actually reported data of confirmed cases [4]. Here, to get C^{av} from the numbers of new infections, we shift forward the infection day of each such case by sampling independently from X and Y.

Given I^{fit}, we proceed as in the illustrative example, but, of course, with "noise" according to these X and Y instead of the gaussians used there. This gives the *reconstructed* numbers of new infections I' (blue curve). Using the infection intensities w_n from [1], we calculate the corresponding reproductive numbers R^{fit} and R'. The latter matches well the green curve R^{TF} that shows the estimated mean reproductive numbers, as published on the group's website [3] on May 7.

The remarks made above on the illustrative example apply also here. We note also that our red curve of reproductive numbers is in good agreement with the results of the inference analysis reported in [5].

The approach by the Robert Koch Institute

Contrary to Switzerland, where, as far as we know, the date of symptom onset of the single cases is not collected systematically, this information is available for the majority of the cases in Germany. For the sake of clarity, we assume here that it is known for all cases.[1] Then, the Robert Koch Institute applies the

[1] We refer to [2] for the method applied to the cases with no known date of symptom onset.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.12.20099366; this version posted May 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

following simpler scheme [2]:[2] The reconstructed infection day i''_a of a case a is the day of its symptom onset minus m, where m is the average incubation time. Counting the number of cases that fall now on day s gives the reconstructed I_s that we denote by I''_s. The reproductive numbers calculated from these I''_s are denoted by R''_s.

Of course we can view the subtracted value m as sample from the distribution of the constant time lag $X = m$. Therefore, this scheme is at least formally very similar to the one adopted by the group at ETH Basel.

From the above discussion it is now clear that the so reconstructed infection times $i''_a = i_a + d''_a$ with "noise" $d''_a = x_a - m$ lead also to a smoothing of the infection number statistics and thus to misleading reproductive numbers. But it is also intuitively clear, that the impact is significantly smaller. This is confirmed by the following calculation. Assume that the above I^{fit} are the true new infections per day and that the distribution of the incubation time and the infection intensity are also as above. Then this scheme gives the following result:

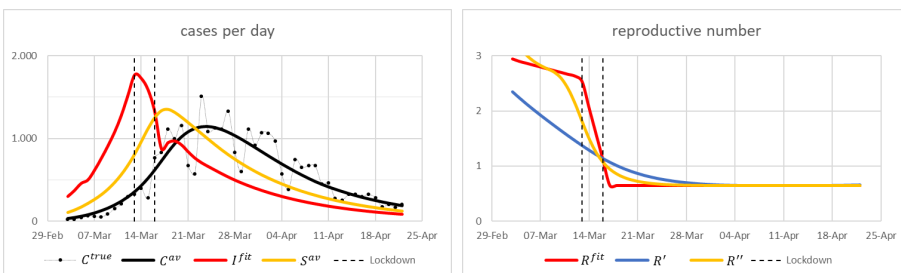

Figure 4: 'True' infection and reproductive numbers (red) and their reconstructions according to the schemes of the group at ETH Basel (blue) and of the RKI (yellow).

We finally remark, that the knowledge of the dates of symptom onset is an advantage, as adopting the same scheme but with dates of confirmation instead of symptom onset, would introduce the "noise" $x_a + y_a - m'$, where m' is the average time between infection and confirmation, into the true data, and this is clearly more "noise" than in the scheme based on dates of symptom onset.

Conclusion

In this note we have reexamined a type of schemes used to estimate the effective reproductive numbers R_t for COVID-19 by the example of two versions actually in use[1, 2]. These schemes are based on reconstruction of not directly observable

[2] As our focus is again on the days around March 17, we restrict our exposition to those days of infection where all infections can be assumed to be confirmed by the actual date of the monitoring. For a method to extend the reconstruction to later days we refer to [2].

medRxiv preprint doi: https://doi.org/10.1101/2020.05.12.20099366; this version posted May 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

data by sampling time lags of observed data from suitable distributions. Noise effects inherent in these schemes smooth the number statistics of the true data. The analysis of thus smoothed number statistics yields stable results and is easier to handle than classical inference methods as applied in [5]. However, under certain circumstances like the current COVID-19 pandemic, the introduced noise effects dominate the information contained in the true data and lead to erroneous interpretations. The simpler approach adopted by the Robert Koch Institute performs better than the one by the group at ETH Basel. Moreover, we point out that adequate knowledge of the date of symptom onset is an advantage.

Acknowledgements

We thank Nicola Kistler for asking the right question and Erik Böttger, Jürg Fröhlich and Emanuel Wyler for helpful discussions.

Disclosure statement

The authors have no financial support nor any other potential conflict of interest relevant to this article.

Corespondence

wyler@physik.uzh.ch

References

[1] Scire J et al. Reproductive number of the COVID-19 epidemic in Switzerland with a focus on the Cantons of Basel-Stadt and Basel-Landschaft. Swiss Medical Weekly 2020;150:w20271 doi: https://doi.org/10.4414/smw.2020.20271

[2] an der Heiden M, Hamouda O. Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland Nowcasting.
Epid Bull 2020;17:1015 — DOI 10.25646/6692.2.
See https://edoc.rki.de/handle/176904/6650.4 for updated version

[3] https://bsse.ethz.ch/cevo/research/sars-cov-2/real-time-monitoring-in-switzerland.html

[4] https://github.com/daenuprobst/covid19-cases-switzerland

[5] Lemaitre J et al. Switzerland COVID-19 Scenarios Report.
https://jcblemai.github.io/