Estimating the SARS-CoV-2 infected population fraction and the infection-to-fatality ratio: a data-driven case study based on Swedish time

N.B. When citing this work, cite the original published paper.

(article starts on next page)

scientific reports

OPEN

Estimating the SARS-CoV-2 infected population fraction and the infection-to-fatality ratio: a data-driven case study based on Swedish time series data

Andreas Wacker[1], Anna Jöud[2], Bo Bernhardsson[3], Philip Gerlee[4], Fredrik Gustafsson[5] & Kristian Soltesz[3]✉

We demonstrate that finite impulse response (FIR) models can be applied to analyze the time evolution of an epidemic with its impact on deaths and healthcare strain. Using time series data for COVID-19-related cases, ICU admissions and deaths from Sweden, the FIR model gives a consistent epidemiological trajectory for a simple delta filter function. This results in a consistent scaling between the time series if appropriate time delays are applied and allows the reconstruction of cases for times before July 2020, when RT-PCR testing was not widely available. Combined with randomized RT-PCR study results, we utilize this approach to estimate the total number of infections in Sweden, and the corresponding infection-to-fatality ratio (IFR), infection-to-case ratio (ICR), and infection-to-ICU admission ratio (IIAR). Our values for IFR, ICR and IIAR are essentially constant over large parts of 2020 in contrast with claims of healthcare adaptation or mutated virus variants importantly affecting these ratios. We observe a diminished IFR in late summer 2020 as well as a strong decline during 2021, following the launch of a nation-wide vaccination program. The total number of infections during 2020 is estimated to 1.3 million, indicating that Sweden was far from herd immunity.

List of symbols

Abbreviations

FIR	Finite impulse response (model)
ICR	Infection-to-case ratio
ICU	Intensive care unit
IFR	Infection-to-fatality ratio
IIAR	Infection-to-ICU admission ratio
LS	Least-squares (method)
RT-PCR	Reverse transcription polymerace chain reaction

Subscripts and modifiers

a	ICU admission
c	Case (detected infection)
d	fatality (death)
$\sim$	denotes raw data before taking averages

[1]Mathematical Physics, Lund University, Lund, Sweden. [2]Occupational and Environmental Medicine, Department of Laboratory Medicine, Lund University, Lund, Sweden. [3]Automatic Control, Lund University, Lund, Sweden. [4]Mathematical Sciences, Chalmers University of Technology, University of Gothenburg, Gothenburg, Sweden. [5]Electrical Engineering, Linköping University, Linköping, Sweden. ✉email: kristian@control.lth.se

Variables and symbols

δ	Dirac's delta distribution
λ	IIAR/ICR
σ	Standard deviation
τ	Time delay (unit: days)
b	Gain parameter
e	Error or residual time series
N	Number of individuals
p	Probability
t	Time index (unit: days)
x	New infection time series (not directly observable)
y	Daily time series (observation)

The COVID-19 pandemic has posed enormous global challenges to the healthcare sector. To estimate the future need of personnel, equipment and hospital beds, reliable statistical analysis tools are required. Historic data is an important asset in figuring out how to best combine available time series data to gain predictive capability while reducing the influence of bias and other sources of prediction error and uncertainty. At the same time, statistical analysis of the historical epidemic evolution can provide indications for the success of medical treatments and vaccination programs. It also allows estimation of the accumulated number of infections. This number essentially determines the level of herd immunity, and thus received much attention in Sweden during the spring of 2020.

It is a difficult task to predict healthcare, and—of particular interest in the COVID-19 context—ICU demand. This is especially true in an early phase of an epidemic caused by a previously unknown pathogen, such as the SARS-CoV-2 virus that causes COVID-19. While it was possible to falsify early prediction models[1] based on high sensitivity, it remains a largely open question how time series data could be analyzed to arrive at accurate and precise predictions, with practical use to healthcare planners.

We investigate if a particular simple class of time-invariant finite impulse response (FIR) models[2]—those with a delayed delta impulse response—is sufficient to model the relation between time series data. Particularly, our aim is to investigate whether the simple FIR model is sufficient for relating COVID-19 cases (detected infections), ICU admissions, and registered deaths in Sweden. We then demonstrate how such simple models can be used for reconstructing the epidemiological evolution during times of measurement uncertainty caused by limited test capacity, as well as prediction of ICU demand based on case data.

Methods

Data used. In this paper, the evolution of the pandemic is based on the following official and openly accessible time series reported by the Swedish Public Health Agency:

- *Cases:* Daily RT-PCR-confirmed SARS-CoV-2 cases in Sweden. The date refers to the registration.
- *ICU admissions:* Daily number of ICU admissions in Sweden for patients with COVID-19 at the given day.
- *Deaths:* Daily number of deaths in Sweden for persons with a SARS-CoV-2 infection at the given day.

The data was downloaded from The Swedish Public Health Agency[3] on 2 November 2021. Here we display (death) data until 10 October 2021 in order to avoid incomplete data due to late reporting. Due to insufficient testing we omit case data before 18 June 2020 in the model fitting, and represent it with dotted lines in the figures. During the first wave in March–May 2020, RT-PCR-testing was essentially focused to persons admitted to hospital and elderly care in Sweden due to limited testing capacity. In the first half of June, the Swedish government strongly advocated the testing of all persons with symptoms of COVID-19 and supplied financial assistance to the regions as of 11 June 2020. We assume that this had full effect on testing after a further week, which justifies the date given above.

In addition to the *ordinary* testing of persons with suspected COVID-19 infection, results for six *randomized* studies[4] in 2020 and 2021 have been published by the Swedish Public Health Agency. They can be used to estimate the prevalence of COVID-19 in the population at the corresponding times. The studies conducted 24–28 August 2020 and 21-25 September 2020 did not provide any positive samples, while 23, 9, 24, and 43 positive cases where detected for 21–24 April 2020, 25–28 May 2020, 30 November–4 December 2020, and 12–16 April 2021, respectively. Test results were available for slightly less than 3 000 persons in the studies of 2020 and 4758 persons for the study in 2021. The limited sample size results in statistical uncertainty, indicated by the 68% confidence limits for the average assuming a Poisson distribution for the number of positively tested. Sampling bias might provide a reduced prevalence for the two latest studies according to the statistical analysis performed in the studies[4]. Here, we use the bare results based on the number of positive cases. For comparison, we also provide an estimate for the ICR based on sampling-bias corrected data.

While deaths and ICU admissions related to COVID-19 naturally also appear in the case data, their total number until 1 November 2021 sums up to only 1.3% and 0.7%, respectively, of the total cases. Regarding ICU and death data, one has to take into account that approximately 75% of ICU patients survive[5], and former ICU patients contribute to the death toll with only 13%, as there are approximately twice as many deaths as ICU admissions. Furthermore, ICU admission data and death data relate to different age groups: While 69% of the ICU patients are younger than 70 years old, 69% of the deceased have reached at least the age of 80. The small overlap between the groups generating the cases, ICU admission, and deaths time series suggests that statistical correlations between the time series can be expected to reflect the links to their common cause: antecedent

SARS-CoV-2 infection in the Swedish society. This motivates the FIR model discussed in "Finite impulse response models" with three independent filter functions for cases, ICU admission, and deaths.

Furthermore, data on antibody prevalence from blood donors and health center samples (unrelated to COVID-19-specific testing) have been aggregated and published[6]. Here we provide 95% confidence intervals for COVID-prevalence based on these data sources. We are aware, that neither the group of blood donors nor the group visiting healthcare centers for other reasons than COVID-19 are representative for the entire population. However, the fact that the prevalence of both groups are very similar, despite their very different characteristics, suggests that their prevalences of antibodies against Covid-19 are representative for the population.

Parameters used. In addition to the data on the COVID-19 evolution and prevalences provided by the Swedish Public Health Agency addressed in "Data used", we apply three further parameters within this study:

- The time interval $T_{\text{interval}} = 10 \pm 1$ days, during which an infected person shows a positive RT-PCR result. See "Calibrating against randomized RT-PCR test data" for details.
- The probability $p_{\text{antibody}} = 0.95 \pm 0.05$, that a previous SARS-CoV-2 infection is detected by an antibody test. See "Comparing with randomized RT-PCR and antibody studies" for details.
- The average duration $\tau_a = 17$ between infection and the admission to ICU. See "Reconstructing the IFR and the IIAR" for details.

Note that T_{interval} and p_{antibody} are used independently of each other in two different determinations of the ICR. Both ways provide essentially the same result, which stabilizes our results against systematic errors in these parameters. τ_a only enters Eq. (22) and the time axis in Fig. 3.

Finite impulse response models. Finite impulse response (FIR) models are a class of linear filters. (The interested reader is referred to the textbook[2] for a thorough mathematical introduction to FIR and related linear model structures.) They describe the outcome of a time-dependent observable (such as a death rate) by a sum of preceding data (here number of infections at earlier dates), which are weighted by a filter function. They are commonly used for analysing epidemiological problems, where the filter function could represent for example the serial interval distribution. For practical applications, the filter function is often not easy to obtain. Here we show that the assumption of a Dirac delta response, where the filter function has only two free parameters (delay and amplitude), allows for a consistent analysis of the COVID-19 evolution in Sweden.

A central entity for the evolution of an epidemic is the number of new infections $\tilde{x}(t)$, occurring on day t. Let $\tilde{p}_c(t, \tau)$ be the probability for an infection starting on day t to generate a reported positive RT-PCR test τ days later. This results in the observation model

$$\tilde{y}_c(t) = \sum_{\tau=0}^{\infty} \tilde{p}_c(t - \tau, \tau)\tilde{x}(t - \tau) + \tilde{e}_c(t), \tag{1}$$

where $\tilde{y}_c(t)$ denotes new cases on day t, and $\tilde{e}_c$ is a zero-mean uncorrelated noise process representing statistical fluctuations associated with the probability $\tilde{p}_c$. The model has finite impulse since p_c is identically zero for sufficiently large τ (e.g. a human lifetime), and can be regarded practically as zero for $\tau \gg 1$ week. The time index t has the unit of days. In (1) it represents that the detection probability distribution, defined through the dependence of the second time index τ, may itself vary over time.

The use of the tilde $\sim$ in (1) is to distinguish unfiltered measurements. Historic observations $\tilde{y}_c(t)$ exhibit a clear weekday pattern. For retrospective analysis it is therefore customary to apply a centered 7-day moving average filter,

$$y_c(t) = \frac{1}{7} \sum_{s=-3}^{3} \tilde{y}_c(t - s), \tag{2}$$

to compensate for such effects. Throughout the paper we will work with time series that have been subjected to filtering according to (2). We will drop the $\sim$ notation but still write e.g. "cases" instead of "filtered cases" in favor of readability.

Within linear system theory, a model with the structure of Eq. (1) is referred to as a (stochastic) finite impulse response (FIR) model, implying (combined with Eq. (2)) that a $y_c(t)$ can be described by a finite record of $x(t)$.

Summation of $p_c(t, \tau)$ over τ, yields the expected infection-to-case ratio (ICR):

$$b_c(t) = \sum_{\tau=0}^{\infty} p_c(t, \tau), \tag{3}$$

defined as the probability that a person infected on day t will eventually become detected and registered as a case.

The central point of the manuscript is that we investigate the hypothesis that $p_c(t, \tau)$ can be adequately modeled using the delta FIR model

$$p_c(t, \tau) = b_c(t)\delta(\tau - \tau_c), \tag{4}$$

where the discrete delta filter function given by

3

$$\delta(t) = \begin{cases} 1, & t = 0, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where τ_c is the average delay between infection and case registration. Note that the model (4) is defined for the averaged quantities (2), where $p_c(t, \tau)$ does not display the weekday fluctuations in t, that are likely in $\tilde{p}_c(t, \tau)$. Assuming that the τ-dependence of $p_c(t, \tau)$ is reasonably well reproduced by its average τ_c and standard deviation σ, the simplified model (Eq. 4) is justified in the Appendix. This relies on the assumption that $p_c(t, \tau)$ does not change on the scale σ in t and that the second derivative of $x(t)$ is much smaller than $x(t)/\sigma^2$. For the special case of an exponential evolution for $x(t)$, this provides an accuracy of better than 5% if σ is less than 46% of the doubling time, as has already been stated[7].

Employing Eq. (4), the observation model Eq. (1) becomes

$$y_c(t) = b_c(t - \tau_c)x(t - \tau_c) + e_c(t). \tag{6}$$

The model (6) asserts that the number of cases $y_c(t)$, detected through RT-PCR testing on day t, only depends on the number of new infections $x(t - \tau_c)$ that occurred τ_c days earlier. Furthermore, the expected dependence is through a linear scaling factor, the ICR.

Relating the time series.   We introduce analogous observation models for ICU admissions y_a and deaths y_d:

$$y_a(t) = b_a(t - \tau_a)x(t - \tau_a) + e_a(t), \tag{7}$$

$$y_d(t) = b_d(t - \tau_d)x(t - \tau_d) + e_d(t), \tag{8}$$

where $b_a(t)$ is the Infection ICU admission Ratio (IIAR) and $b_d(t)$ the Infection Fatality Ratio (IFR), where the time dependence denotes the infection date.

The underlying infections $x(t)$ are unknown, and cannot be estimated solely from the measurements y_c, y_a, y_d, since an absolute reference frame against which to estimate the individual time-shifts and gain factors is not available. However, if we disregard the noise terms, we can relate the cases and ICU admissions time series through

$$y_a(t) = \frac{b_a(t - \tau_a)}{b_c(t - \tau_a)} y_c(t - \tau_{ac}), \tag{9}$$

where $\tau_{ac} = \tau_a - \tau_c$ is the average delay between the registration as a case and the admission to ICU (not necessarily for the same person). Analogously, we have

$$y_d(t) = \frac{b_d(t - \tau_d)}{b_c(t - \tau_d)} y_c(t - \tau_{dc}), \tag{10}$$

with $\tau_{dc} = \tau_d - \tau_c$ and

$$y_d(t) = \frac{b_d(t - \tau_d)}{b_a(t - \tau_d)} y_a(t - \tau_{da}), \tag{11}$$

with $\tau_{da} = \tau_d - \tau_a$.

Equations (9–11) can be conveniently fitted to data. If the time-dependence of the b-coefficients is negligible, we can fit the ratio $\lambda = b_a/b_c$ and τ_{ac} from Eq. (9) by minimising the sum of squares

$$F_{ac}^{\text{LS}}(\lambda, \tau_{ac}) = \sum_t \left[y_a(t) - \lambda y_c(t - \tau_{ac}) \right]^2, \tag{12}$$

which results in the two fitting parameters λ, τ_{ac}. In order to obtain robust estimates we alternatively minimise the modified sum of squares

$$F_{ac}^{\text{LSmod}}(\lambda, \tau_{ac}) = \sum_t \left[\frac{y_a(t)}{\sqrt{\lambda}} - \sqrt{\lambda} y_c(t - \tau_{ac}) \right]^2. \tag{13}$$

As a third option, we maximize the correlation coefficient

$$r(\tau_{ac}) = \frac{\sum_t \left[y_a(t) - \bar{y}_a \right] \left[y_c(t - \tau_{ac}) - \bar{y}_c \right]}{\sqrt{\sum_t \left[y_a(t) - \bar{y}_a \right]^2} \sqrt{\left[y_c(t - \tau_{ac}) - \bar{y}_c \right]^2}} \tag{14}$$

to obtain τ_{ac} and use $b_a/b_c = \bar{y}_a/\bar{y}_c$. In all three approaches, the times t are chosen such that reliable data for both $y_c(t - \tau_{ac})$ and $y_a(t)$ are available. Equations (10, 11) are treated in the same way, where the indices a, c are replaced by d, c and d, a, respectively, in the formulae above. Note, that each combination of indices applies a different time interval due to the reliability condition.

Calibrating against randomized RT-PCR test data.   While Eqs. (9–11) establish relative relations between the time series y_c, y_a, y_d, a "grounding point" is needed to obtain absolute values of the time shift and

scaling parameters of the observation models Eqs. (6–8). Randomized RT-PCR studies provide such a grounding point, where we use the data discussed in "Data used". From the number of positive RT-PCR test results in each study, we can estimate the prevalence $N_{\text{positive}}(t)$ by multiplying with the population of Sweden and dividing by the number of tested persons.

The prevalence $N_{\text{positive}}(t)$ depends on the probability $p_{\text{positive}}(\tau)$ to have a positive test result τ days after becoming infected:

$$\tilde{N}_{\text{positive}}(t) = \sum_{\tau} p_{\text{positive}}(\tau)\tilde{x}(t - \tau). \tag{15}$$

After time-averaging and using again the impulse FIR model this provides

$$N_{\text{positive}}(t) = T_{\text{Interval}}x(t - \tau_{\text{positive}}), \tag{16}$$

where

$$T_{\text{Interval}} = \sum_{\tau} p_{\text{positive}}(\tau) \tag{17}$$

is the average time-interval over which a positive test result is expected and

$$\tau_{\text{positive}} = \frac{1}{T_{\text{Interval}}} \sum_{\tau} \tau p_{\text{positive}}(\tau) \tag{18}$$

is the average delay after the time of infection. From the data displayed in Figure 2 of Kuricka et al.[8], we extract $T_{\text{Interval}} = 10.8$ days and $\tau_{\text{positive}} = 12$ days. Another study[9] found $T_{\text{Interval}} = 9.5$ days. Motivated by these numbers we use $T_{\text{Interval}} = 10 \pm 1$ days and make the simplifying assumption $\tau_{\text{positive}} = \tau_c$. (The resulting value for τ_c from Eq. (22) agrees well with $\tau_{\text{positive}} = 12$ days, extracted from[8]). Analogously to Eq. (9) we fit the relation

$$N_{\text{positive}}(t) = \frac{T_{\text{Interval}}}{b_j(t - \tau_{\text{positive}})}y_j(t - \tau_{\text{positive}} + \tau_j) \tag{19}$$

where $j = a, d$ refers to the data sets for ICU and deaths. We apply all three fitting routines, again neglecting the time-dependence of $b_j(t)$—however, $\tau_{\text{positive}} = \tau_c$ is kept fixed. (As we only have 4 non-vanishing data points for N_{positive}, the use of a second fit-parameter next to b_j could provide spurious results)

Results
Scaling of data.
The symbols on the upper panel of Fig. 1 show the daily Swedish numbers of positively tested persons (named cases), persons admitted to intensive care units (named ICU), and deaths, see "Data used" for details. These data were averaged over a seven day period (lines) in order to avoid weekday fluctuations, resulting in the functions $y_c(t)$ (Cases), $y_a(t)$ (ICU admission), and $y_d(t)$ (deaths), which are the main data-sets used throughout this article. Here the time t is chosen as the central day of the averaging period.

As described in "Relating the time series", we determined the ratios b_a/b_c, b_d/b_c, and b_d/b_a, as well as the corresponding delays. The vaccination program against SARS-CoV-2 (starting in the last week of 2020 in Sweden) is likely to have effected $b_c(t)$, $b_d(t)$ and $b_a(t)$ differently. In its first phase, the program targeted elderly in need of special care, who show the by far largest risk to die from COVID-19. Thus, we disregard death data after January 2021 for the model fitting. Subsequently, groups were vaccinated in order of decreasing age. The main part of the age group 65–70 years was vaccinated with the first dose around the first half of April 2021. As this group is highly over-represented in ICU care, we also disregard ICU admission data newer than April 2021 from the model fitting. As mentioned above, we additionally disregard the case data before 18 June 2020, when testing became available to all persons with symptoms in Sweden, see "Data used". The results are given in Table 1. We find that all three approaches provide identical time delays, which we regard as particularly reliable. Also the fractions between the ratios agree fairly well with deviations far below 10 %. In the following we apply the values from the least squares (LS) method, see "Relating the time series", but we note that none of our results depends on this choice of method. The relations $b_a/b_c \cdot b_d/b_a = b_d/b_c$ and $\tau_{ac} + \tau_{da} = \tau_{dc}$ hold only approximately as different time intervals are used in the fitting due to the exclusion of death data after 1 February 2021 and case data before 18 June 2020 addressed above.

Using these scaling factors $b_a/b_c = 1/193$ and $b_d/b_c = 1/68$ as well as the respective time delays, the lower panel of Fig. 1 shows that all three curves agree very well over the second wave of November 2020–January 2021. The ICU and death curve also show a similar behavior at the first wave March–May 2020, albeit the ratio between ICU admission and deaths appears to be slightly higher here. The case numbers are much lower due to the limited testing before mid-July. For the third wave March–May 2021, the ICU and case curve agree very well (the dip in cases around 1 April may be attributed to decreased testing around Easter). We also see that the death curve shows much lower values from around mid-January 2021, which coincides with the start of the vaccination program in Sweden at the end of 2020.

Comparing with randomized RT-PCR and antibody studies.
We also fitted the data from the 6 randomized RT-PCR studies to the ICU and death data (here we omitted the sixth study, as the fatality was significantly reduced in 2021, most likely due to the vaccination program) and found almost identical results in both cases, see Table 1. Thus, the prevalence of RT-PCR-detectable SARS-CoV2 infections is about 16 times higher
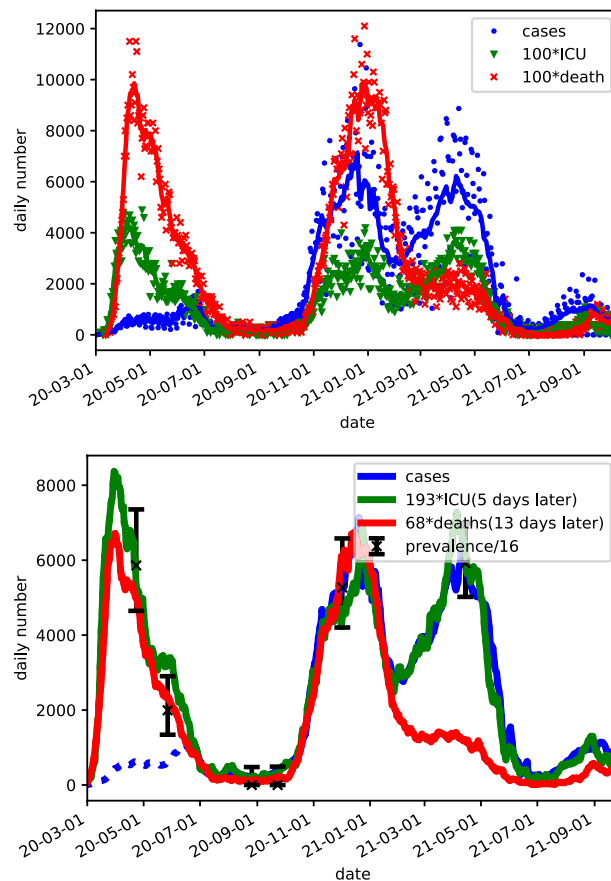
Figure 1. Upper panel: Raw data (symbols) used in this study. (The numbers of ICU admissions and deaths have been multiplied by 100 to provide comparable numbers.) The lines show the respective averages over a 7 day period, where the central day of the interval was used in the abscissa. Lower panel: 7 day averaged data from upper panel scaled using the proposed methodology of "Methods", with resulting scale factors provided in the figure legend. The ICU-data are shifted by 5 days to the left and the death-data are shifted by 13 days to the left, so that they essentially fall on one curve. Dotted lines indicate sections of data that are considered as incomplete. The black error bars provide scaled 68% (one normal standard deviation) confidence levels of randomized RT-PCR study data. All scaling factors and delays are taken from the LS values in Table 1.

Parameter	LS	LSmod	Corr. coeff.
b_a/b_c	0.0052	0.0051	0.0051
τ_{ac}	5	5	5
b_d/b_c	0.0147	0.0146	0.0141
τ_{dc}	13	13	13
b_d/b_a	2.70	2.65	2.59
τ_{da}	8	8	8
$\frac{T_{\text{Interval}}}{b_a}\frac{b_a}{b_c}$	15.9	16.0	15.2
$\frac{T_{\text{Interval}}}{b_d}\frac{b_d}{b_c}$	15.8	15.9	14.7

Table 1. Results for different fitting procedures (*LS* least squares; *LSmod* modified least squares; *Corr. coeff.* correlation coefficient) detailed in "Relating the time series" for ratios and time delays. In the last two rows the fitted values for T_{Interval}/b_a and T_{Interval}/b_d were multiplied with the factors from row 1 and 3 respectively to obtain estimates for T_{Interval}/b_c. See List of symbols at the begining of the article for an explanation of the parameters.
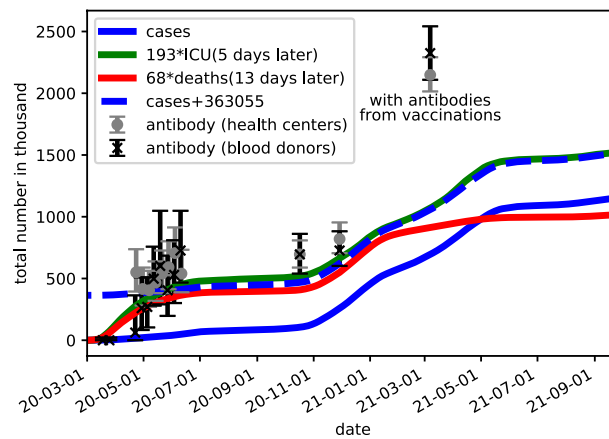
Figure 2. Accumulated number of cases at a given date, where we added the estimates based on scaling the death and ICU data. The dashed blue line provides the case data where we added an estimate for number of cases missed due to limited testing before mid-June. The symbols with error bars show the results of antibody tests performed for blood donors and blood samples from health centers[6], where the fraction of positive tests was multiplied by Sweden's population. Note that the number of persons with antibodies in March 2021 include detected results from vaccinations.

than the number of cases (as reconstructed by shifting and scaling the death and ICU data), see the lower panel of Fig. 1. As an infected person can be detected over an average time interval T_{interval}, but is only registered once as a case with probability b_c, this implies $T_{\text{Interval}}/b_c \approx 15.6 \pm 0.9$. See "Calibrating against randomized RT-PCR test data" for details. Here we used the average of all values provided in the two lowest rows of Table 1 and used the maximal deviation as an estimate for the error. The time interval, an infected person is tested positive seems to be less well known. Using $T_{\text{Interval}} = 10 \pm 1$ days, see "See "Calibrating against randomized RT-PCR test data"", this provides the ICR based on the prevalence studies

$$b_c^{\text{prevalence}} \approx 0.64 \pm 0.10 \tag{20}$$

We note that the sampling-bias corrected data for the prevalence results in a slightly larger value $b_c^{\text{prevalence\_corrected}} \approx 0.7 \pm 0.11$ with overlapping error marginal.

The lower panel of Fig. 1 shows, that the case data agree with the prevalences divided by 16 after testing become widely accessible around mid-June 2020. As both the scaled ICU and death data agree well with the prevalences at all times, we can take these curves for estimating the number of cases before mid-July, where the case data are not reliable due to limited testing.

In Fig. 2 we find a clear plateau of total cases in July–November 2020, as there were few new cases in this period. The average plateau value (at 1. Sept) was 450,000 (reconstructed) cases, with an uncertainty (based on the ICU and death data) of about 55,000. Antibody tests for different groups provide estimates of 700,000 positive persons in Sweden both at the beginning and the end of the plateau, see Fig. 2. This results in $b_c/p_{\text{antibody}} = 0.64 \pm 0.08$. The probability p_{antibody} to develop detectable antibodies has been found[10,11] to be above 90%. As it should not exceed 100%, we assume $p_{\text{antibody}} = 0.95 \pm 0.05$, and find

$$b_c^{\text{antibody}} \approx 0.61 \pm 0.11 \tag{21}$$

This agrees very well with the different estimate Eq. (20), albeit different statistical and systematic errors (in particular the values of T_{Interval} and p_{antibody}) enter both ways to calculate b_c. Thus we consider $b_c = 0.63$ as a good estimate for the ICR with the awareness that a 10% error is not unlikely.

We note, that a relatively large number of persons with antibodies was found in the study at the beginning of March 2021. Here, one has to take into account that a part of the antibodies detected results from vaccinations. At this time about 700,000 persons had been vaccinated in Sweden and a majority of them should have developed antibodies, when the data was collected.

Reconstructing the IFR and the IIAR. Assuming the ICR $b_c = 0.63$ for the time after June 18th, 2020, we can estimate the variables $b_a(t)$ and $b_d(t)$ from Eqs. (9, 10). At first we need the absolute delays. Here we rely on the data for ICU admission, which in average occurs about 11 days after the onset of symptoms according to the Swedish Intensive Care Registry[12]. Furthermore, it is known, that it takes about 6 days from the times of infection to develop symptoms[13–15]. Thus we use $\tau_a = 17$ days in the following. Based on the values on Table 1 we get

$$\tau_c = 12 \ \text{days}, \tau_a = 17 \ \text{days, and } \tau_d = 25 \ \text{days.} \tag{22}$$

In Fig. 3 we plot the time-dependence of the infection-to-ICU admission ratio (IIAR) $b_a(t)$ by a green line on the basis of Eq. (9). Here we used 21 day averages to restrict fluctuations. We find that $b_a(t) \approx 0.34$ % is close
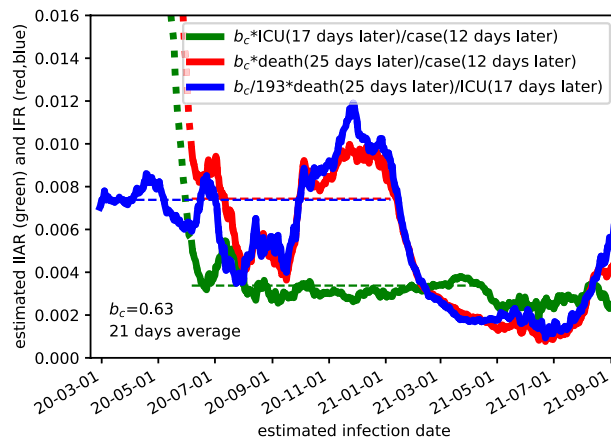
Figure 3. Estimated values for the IIAR $b_a(t)$ (green) and the IFR $b_d(t)$ (red, blue) based on Eqs. (9, 10, 11). We assume a constant ICR $b_c = 0.63$, which appears reliable for cases after 18 June (dotted curves indicate that earlier cases data are applied). The horizontal dashed lines provide the average values.

to its average, which confirms the quality of the scaling. The larger bump around late July occurs within a range with small numbers of ICU admissions (average of 1.6 in August 2020), so that statistical fluctuations cannot be excluded here.

Similarly, we obtain the IFR $b_d(t)$ from Eq. (10) as shown by the red line in Fig. 3. On average, we have 0.74% until end of 2020 (before vaccinations started to show effects), but there are pronounced changes over time. For infections in August and September of 2020, the IFR is much lower. For infections after 1 January 2021, we see a distinct decline in the IFR reaching values of 0.1% for infections around mid-June 2021. Afterwards the IFR predicted by our approach increased again. The same behavior for the IFR can be obtained from from Eq. (11) as shown by the blue line in Fig. 3, which now extrapolates to times before mid-June under the assumption, that the IIAR remained essentially constant in this period as well.

Discussion

Using openly accessible data from the Swedish Public Health Agency and the Swedish Intensive Care Registry, our approach provides estimates for the infection-to-fatality ratio (IFR), infection-to-case ratio (ICR), and infection-to-ICU admission ratio (IIAR). We find that data for daily cases, daily ICU admission, and daily deaths of individuals with confirmed infection fall on essentially a single curve based on a FIR model with a delta filter function and time-invariant fit parameters, see the lower panel of Fig. 1. There are only two major well-understood exceptions: (I) Cases exhibit a poor match before mid-June 2020, when free RT-PCR testing became broadly available in Sweden for all persons with symptoms. (II) There is a sharp relative decrease in deaths coinciding with the start of the national vaccination program around the turn of the year 2020–2021.

Based on these findings we demonstrate that the approach can be used to retrospectively estimate the cases time series prior to July 2020, that would have been observable with a broad RT-PCR testing program in place. Furthermore, by incorporating data from six randomized RT-PCR studies conducted by the Swedish Public Health Agency and data for the prevalence of antibodies, we obtain an absolute value for the infection-to-case ratio ICR of 0.63 ± 0.07 for the time when testing is easily available to persons with symptoms in Sweden. This value is larger (but compatible within its error) than the value 0.56 obtained in a study for Iceland[10]. It means that approximately 37% of all infected remained undetected as cases.

The key underlying assumption is that the correlation of the data observed under the model originates in causal relations. Such assumption is present, and constitutes a limitation, in all dynamic modeling. The reason we found merit in the proposed model is that it (I) is supported by data, (II) has few parameters, thus reducing the concern of over-fitting, and (III) is consistent with the plausible mechanistic explanation presented in the manuscript.

Additionally, the model relies on the assumption that the confusion matrices (true and false positives and negatives) for involved time series did not change importantly over time. For example, we have assumed that the RT-PCR detection probability is similar across SARS-CoV-2 variants (but it is straightforward to replace this assumption with externally obtained information on how e.g. detection probabilities vary based on the exact testing protocol). This is why we have omitted data from episodes during which such variations are known to have been prominent: early in the pandemic, when there was a shortage of RT-PCR test availability; late in the pandemic, when the effect of vaccination started to manifest itself in the data. For the period in-between, the good correlation of data under the model suggests that the assumption is a valid one, or at least one that cannot be rejected by the considered data.

Regarding systematic errors, the ICR value obtained from our study is reduced if both the time interval for positive testing, T_{Interval}, and probability to develop measurable antibodies after an infection, p_{antibody}, turn out to be much less than 10 days and 0.95, respectively. On the other hand, larger values are less likely as $p_{\text{antibody}} = 0.95$ was chosen close to its absolute maximum value 1 in the analysis. However, a sampling bias or false-positive tests may allow for deviations, which we cannot judge here.

Figure 2 indicates that the total number of cases would have been around 820 000 at the end of 2020, if testing in the first half of 2020 had been as available as in the second half. With an ICR of 0.63, this provides about 1.3 million infected persons in Sweden in 2020. This corresponds to 13% of the population and is far below values required to reach herd immunity.

Based on the ICR of 0.63, we obtain the infection to ICU admission ratio IIAR of 0.34% and an average infection to fatality ratio IFR of 0.74 % (before the start of the vaccinations), see Fig. 3. Our value for the IFR is comparable to earlier studies[16–18]. Note that possible systematic errors in ICR, as addressed above, affect the IIAR and IFR proportionally. Our analysis shows that the IIAR is rather constant in time, at least for the time after august 2020. In contrast, the IFR varies much stronger over time. We attribute its decline starting for infections around 1 January 2021 to successful vaccination of elderly persons. They dominate the mortality in COVID-19, but are less relevant for ICU admission. The increase in the IFR after early July 2021 may have a variety of reasons. It may indicate an increasing number of breakthrough infections for the elderly around 6 months after their vaccination. However, another contributing factor could be the fact that around 1 July 2021 the SARS-CoV-2 vaccination became available to all adults in Sweden. In Sweden, the following vaccines are approved (approval date): Comirnaty by Pfizer/BionTech (21 December 2020), Spikevax by Moderna (January 6 2021), Vaxzevria by Astra Zenecas (29 January 2021), and Covid-19 Cavvine Janssen by Janssen (11 March 2021). Predominantly, Comirnaty has been used, making up for roughly 13 M of totally 19 M doses administered according to Swedish Public Health Agency statistics[19] by 29 October 2021.

The appearance of novel variants of the virus in Sweden during late 2020 and 2021 complicates the interpretation of the results. The disease severity is known to vary depending on the variant[20], but the time period during which new variants appeared overlaps with the time frame of the vaccination program, making it difficult to disentangle the effects of vaccination and new variants on the IFR. Relatedly, we lack explanation for the reduced IFR between mid-July and mid-September of 2020, as seen in Fig. 3. Such observations suggest that modeling for prevalence and healthcare demand estimation or prediction could be facilitated by national-level aggregation of epidemiological data. Particularly, accessible data on vaccination status, relative prevalence of identified virus variants, and reasons (suspected infection, randomized testing, etc.) could be directly incorporated into the proposed model with the objective of bias reduction.

The close correlation between cases, ICU admissions, and deaths—down to a linear scaling and time-shift once self-test had been made broadly available—warrants further investigation. It could be explained through (I) time-invariance of the ICR, the IIAR, the IFR, (the dip in IFR in summer is not visible in the lower panel of Fig. 1 due to the small numbers) and the temporal distributions describing the associated flows; (II) variations in the mentioned entities that have essentially cancelled each other throughout 2020; (III) external confounders providing this canceling effect. Case (I) would imply that the healthcare system's ability to save COVID-19 patients and the impact from different virus mutations has not changed markedly. Case (II) would be surprising for the time from August 2020, where the ratio between cases and ICU admission is largely constant in time, see Fig. 3. Thus it is most likely that $b_c(t)$ and $b_a(t)$ are both constant in this range. However, for the first half of 2020, things are less obvious. The variations observed in the magenta curve may result from reductions in $b_a(t)$ and $b_d(t)$ (i.e. improvement in healthcare) occurring at different times. Case (III) would also be noteworthy since the groups of deceased and persons admitted to ICU care have marginal overlap with each other.

The studied time series alone are not sufficient to map out the causation of the observed correlations, thus distinguishing between (I) and combinations of (II) and (III). However, an understanding of the underlying mechanism could be obtained by retrospectively tracking individual traces connecting the considered time series (e.g. persons who have tested positive, been admitted to an ICU or died with COVID-19). Importantly, if the correlation can be understood and shown not to be coincidental, it could constitute the basis for an accurate 1–2 week predictor of the ICU demand.

Conclusion

The summarizing conclusion from our observations is that important insight into numerous aspects of an ongoing epidemic can be obtained by considering the scaling between different time series, where time shifts are crucial. This is based on an FIR model with a delta filter function, which is shown to work well. We demonstrated this by reconstructing the daily number of cases for the first half year of 2020 periods, where testing was limited in Sweden, and extracting time variations in the infection fatality ratio.

Appendix: Quality of the delta-impulse model

After averaging to remove weekly fluctuations and neglecting fluctuations $e_c(t)$, (1) provides

$$y_c(t) = \sum_{\tau=0}^{\infty} p_c(t - \tau, \tau)x(t - \tau). \tag{23}$$

While we do not know much about $p_c(t, \tau)$, we make the reasonable assumption that it has a single peak and does decay rather quick for large delays τ. In this case it is reasonably represented by its average $\bar{\tau}$ and standard deviation σ defined as

$$\bar{\tau}(t) = \frac{1}{b_c(t)} \sum_{\tau=0}^{n} \tau p_c(t, \tau), \tag{24}$$

$$\sigma^2(t) = \frac{1}{b_c(t)} \sum_{\tau=0}^{n} (\tau - \bar{\tau}(t))^2 p_c(t, \tau), \tag{25}$$

where we used the normalisation from Eq. (3). Assuming, that $x(t)$ is a smooth function, which appears reasonable after 7-days averaging, we may approximate by a second order Taylor expansion

$$x(t - \tau) \approx x(t - \tau_c) - x'(t - \tau_c)(\tau - \tau_c) + \frac{1}{2} x''(t - \tau_c)(\tau - \tau_c)^2, \tag{26}$$

where, for given t, the average delay τ_c satisfies $\tau_c = \bar{\tau}(t - \tau_c)$. Then we obtain from Eq. (23) that

$$y_c(t) \approx \sum_{\tau=0}^{\infty} p_c(t - \tau, \tau) \left[x(t - \tau_c) - x'(t - \tau_c)(\tau - \tau_c) + \frac{1}{2} x''(t - \tau_c)(\tau - \tau_c)^2 \right]. \tag{27}$$

Assuming $p_c(t - \tau, \tau) \approx p_c(t - \tau_c, \tau)$, as the detection probability should not change within a time-scale of σ for constant delay τ, we find

$$y_c(t) \approx b_c(t - \tau_c) x(t - \tau_c) + \frac{\sigma(t - \tau_c)^2}{2} b_c(t - \tau_c) x''(t - \tau_c). \tag{28}$$

The first term is just our delta-impulse model, see Eq. (4), while the second term provides a correction less than 5% if

$$\sigma(t - \tau_c)^2 \frac{|x''(t - \tau_c)|}{x(t - \tau_c)} < 0.1. \tag{29}$$

If the infections show an exponential behavior $x(t) = x_0 e^{rt}$, this provides $\sigma r < 0.316$ or $\sigma < 0.46 t_{\text{doubling}}$ with the doubling time $t_{\text{doubling}} = \ln 2/r$. The last relation has been previously published[7] for the applicability of the delta-response for exponential evolutions. Here we generalised this to arbitrary evolutions $x(t)$, where $t_{\text{doubling}}/\ln 2$ is replaced by the square root of the inverse relative second derivative.

References

1. Soltesz, K. *et al.* The effect of interventions on COVID-19. *Nature* **588**, E26–E28. https://doi.org/10.1038/s41586-020-2405-7 (2020).
2. Antsaklis, P. & Michel, A. *Linear Systems* (Birkhauser, 2005).
3. Swedish Public Health Agency. Bekräftade fall i sverige - daglig uppdatering (2021). https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/bekraftade-fall-i-sverige/. Accessed 01 Nov 2021.
4. Swedish Public Health Agency. Undersökningar avseende förekomst av sjukdom (2021). https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/undersokningar-och-datainsamlingar/forekomst-av-sjukdom/. Accessed 01 Nov 2021.
5. Strålin, K. *et al.* Mortality trends among hospitalised COVID-19 patients in Sweden: A nationwide observational cohort study. Lancet Region. Health Eur. **4**, 100054. https://doi.org/10.1016/j.lanepe.2021.100054 (2021) .
6. Swedish Public Health Agency. Undersökningar avseende genomgången infektion (2021). https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/undersokningar-och-datainsamlingar/genomgangen-infektion/. Accessed 01 Nov 2021.
7. Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B Biol. Sci.* **274**, 599–604. https://doi.org/10.1098/rspb.2006.3754 (2007).
8. Kucirka, L., Lauer, S., Laeyendecker, O., Boon, D. & Lessler, J. Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time since exposure. *Ann. Intern. Med.* **173**, 262–267. https://doi.org/10.7326/M20-1495 (2020).
9. Hu, Z. *et al.* Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. *Sci. China Life Sci.* **63**, 706–711. https://doi.org/10.1007/s11427-020-1661-4 (2020).
10. Gudbjartsson, D. *et al.* Humoral immune response to SARS-CoV-2 in Iceland. *N. Engl. J. Med.* **383**, 1724–1734. https://doi.org/10.1056/NEJMoa2026116 (2020).
11. Dan, J. *et al.* Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science* **371**. https://doi.org/10.1126/science.abf4063 (2021).
12. The Swedish Intensive Care Registry (2021). https://www.icuregswe.org/data--resultat/COVID-19-i-svensk-intensivvard. Accessed 01 Nov 2021.
13. McAloon, C. *et al.* Incubation period of COVID-19: A rapid systematic review and meta-analysis of observational research. *BMJ Open* **10**. https://doi.org/10.1136/bmjopen-2020-039652 (2020).
14. Lauer, S. *et al.* The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **172**, 577–582. https://doi.org/10.7326/M20-0504 (2020).
15. Qin, J. *et al.* Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. *Sci. Adv.* **6**. https://doi.org/10.1126/sciadv.abc1202 (2020).
16. Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: A model-based analysis. *Lancet Infect. Dis.* **20**, 669–677. https://doi.org/10.1016/S1473-3099(20)30243-7 (2020).
17. Levin, A. *et al.* Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *Eur. J. Epidemiol.* **35**, 1123–1138. https://doi.org/10.1007/s10654-020-00698-1 (2020).
18. Swedish Public Health Agency. The infection fatality rate of COVID-19 in Stockholm (2020). https://www.folkhalsomyndigheten.se/contentassets/53c0dc391be54f5d959ead9131edb771/infection-fatality-rate-covid-19-stockholm-technical-report.pdf. Accessed 01 Nov 2021.

19. Swedish Public Health Agency. Leveranser av vaccin mot covid-19 (2021). https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/leveranser-av-vaccin/. Accessed 04 Nov 2021.
20. Twohig, K. A. *et al.* Hospital admission and emergency care attendance risk for sars-cov-2 delta (b. 1.617. 2) compared with alpha (b. 1.1. 7) variants of concern: A cohort study. *Lancet Infect. Dis.* (2021).

Acknowledgements

Author contributions

The study was initiated by A.W. who also produced all figures. K.S., F.G., and B.B. suggested and elaborated the technical methods used. A.J. and P.G. provided the epidemiological perspective. All authors contributed essentially to the writing of the manuscript and approved the final version.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.