Suhail et al. BMC Medical Research Methodology (2021) 21:11
https://doi.org/10.1186/s12874-020-01196-4

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Incorporating and addressing testing bias within estimates of epidemic dynamics for SARS-CoV-2

Yasir Suhail[1,2*], Junaid Afzal[3] and Kshitiz[1,2*]

Abstract

Background: The disease burden of SARS-CoV-2 as measured by tests from various localities, and at different time points present varying estimates of infection and fatality rates. Models based on these acquired data may suffer from systematic errors and large estimation variances due to the biases associated with testing. An unbiased randomized testing to estimate the true fatality rate is still missing.

Methods: Here, we characterize the effect of incidental sampling bias in the estimation of epidemic dynamics. Towards this, we explicitly modeled for sampling bias in an augmented compartment model to predict epidemic dynamics. We further calculate the bias from differences in disease prediction from biased, and randomized sampling, proposing a strategy to obtain unbiased estimates.

Results: Our simulations demonstrate that sampling biases in favor of patients with higher disease manifestation could significantly affect direct estimates of infection and fatality rates calculated from the numbers of confirmed cases and deaths, and serological testing can partially mitigate these biased estimates.

Conclusions: The augmented compartmental model allows the explicit modeling of different testing policies and their effects on disease estimates. Our calculations for the dependence of expected confidence on a randomized sample sizes, show that relatively small sample sizes can provide statistically significant estimates for SARS-CoV-2 related death rates.

Keywords: SARS-CoV-2, Epidemiology, Sampling bias, Covid-19, inaccurate epidemic predictions, overestimation of COVID death rate

* Correspondence: yasir.suhail@uconn.edu; kshitiz@uchc.edu
[1]Department of Biomedical Engineering, University of Connecticut Health, Farmington, CT, USA
Full list of author information is available at the end of the article

Background

The spread of SARS-CoV-2 across the world has led to a significant disease burden with widespread health impact. Optimally planning and implementing such interventions is intimately connected with epidemiological disease modeling and requires the estimation of key metrics such as the speed of infection spread, recovery and fatality rates, and kinetics related to the persistence or loss of acquired immunity. Early reports from the World Health Organization (WHO) stated a case fatality rate of over 3.8% [1] for SARS-COV-2 as it was first detected in Wuhan, China and spread across the world. Subsequently, epidemiological modeling and projection, with its inherent estimation of infection, recovery, and fatality rates has become central to various institutional actors dealing with the management of the epidemic. These studies and reports are, by necessity, ultimately based on the reported numbers from tested patients, which were sampled by public health agencies in the countries where the virus had started spreading [2–5]. However, there are wide differences across countries in the number of people who were tested, the availability of test kits, as well as the stratification of the population that were tested. Due to these factors, the infection [6–8], and case fatality rates (CFR) based on current SARS-CoV-2 data are difficult to interpret [2–4], masking the true extent and dynamics of the disease spread and ensuing fatality.

Caution has been raised with the initial spread of the pandemic, including by us, regarding the accuracy of determined fatality rates [9, 10]. The accuracy of the fatality estimates could be dependent on two potentially important issues addressed in the manuscript: (1) underlying spread of immunity within the population, and (2) ascertainment bias. Reverse transcription polymerase chain reaction (RT-PCR), the commonly employed method to confirm the presence of SARS-CoV-2, only informs about the live status of the virus in the population, and therefore masks the percentage of people who contracted the virus and subsequently resolved the infection by acquired immunity [11]. SARS-CoV-2 is known to induce a detectable antibody response following a few days of infection [11–13], with evidence that this acquired immunity may also be affected by the prevalence of other respiratory viruses [14]. Furthermore, even the antibody response is expected to last temporarily, before memory T cells are formed and archived, which could clonally expand and mount a response for similar subsequent infections [15–17]. An initial report suggested that a larger cohort of tested populations which were negative for an active viral load can now be regarded as having previously contracted the virus [13], but the effect size found was smaller than could be statistically determined from the error rate of the underlying test. Apart from a few studies [18, 19], the true fatality rate of SARS-CoV-2 has not been established, either in general population or in a stratified subpopulations. This in turn has grave implications for public health capacity planning and intervention decisions beyond the next few months.

Secondly, we highlight the bias within the sampling (testing for SARS-CoV-2 presence), which could potentially alter the estimates of both the infection and fatality rates. There have been multiple mathematical studies modeling the kinetics of disease spreading with and without social distancing interventions [20–23]. However, these are dependent on model parameters estimated from limited, and likely biased and non-uniform sampling [24–26]. Indeed, these numbers vary widely across different countries, resulting in large variations in suggested mortality rates [23, 27]. Incidentally, these data were collated for the objective of public health operations, identifying infected individuals and tracing their contacts etc., and for preparation of adequate health facilities. However, these approaches may introduce incidental bias in testing for individuals presenting with symptoms, rendering models built on these data vulnerable to systematic sampling bias. This raises significant concerns regarding the accuracy of the estimates of fatality and morbidity rates, with far reaching consequences on capacity planning and policy making. Although, kinetic modeling studies have attempted to mitigate or sidestep the effect of sampling bias by various methods. For example, Verity et al. give estimates for the infection fatality rates based on testing of foreign nationals repatriated from China [2]. While this sample may not have been directly biased with symptom severity, it is still likely to be highly correlated to age, health, and placement within social and physical contact networks, and therefore indirectly correlated with infection status and susceptibility to fatality. Therefore, it is essential to understand the effect of the sampling bias in prediction of fatality rates where sampling bias cannot be avoided (e.g. in hospitals preferentially testing symptomatic or more severely affected patients), as well as to use unbiased sampling to ascertain the true fatality rates by public policy authorities.

In this work, we (i) present a new model of the epidemic dynamics by augmenting the commonly employed SIRD compartment model (with four distinct population stratification: S: susceptible, I: infected, R: recovered, D: dead) and explicitly introducing bias in their sampling; (ii) demonstrate by model simulation that testing bias could have

significant effect on the estimation of the infection and fatality rates; and finally (iii) propose that an unbiased randomly sampled testing study in a region with high current fatality presents the best course to estimate the true fatality rates. Similar treatment of sampling bias was also presented by Brunner and Chia [28], concentrating only on estimates of disease spreading. Using our augmented SIRD model with incorporation of serological testing, we show that estimation of acquired immunity could partially mitigate the effect of testing bias, and demonstrate that case fatality rates may underestimate infection fatality rates because of the lag between infection and death. Our calculations indicate that a reasonable unbiased testing sample can provide high confidence data to test the hypotheses of different fatality rates.

Together with our augmented compartmental model, our proposed scheme presents a coherent, statistically rigorous estimation method to determine infection and fatality rates, which is both cognizant of the testing bias in favor of the more symptomatic or severe patients, and given sufficient follow-up time, resistant to the underestimation bias due to lagging death counts.

Methods

SIRD model with disease stratification into high and low symptomatic populations

First, we augment the canonical SIRD model by stratifying the infected population into high (H) and low (L) symptom populations. The differential equations for disease progression can be written as

$$\begin{aligned}
\frac{dS}{dt} &= -\beta\frac{S(H+L)}{N} \\
\frac{dH}{dt} &= \beta q\frac{S(H+L)}{N} - \gamma I \\
\frac{dL}{dt} &= \beta(1-q)\frac{S(H+L)}{N} - \gamma L \\
\frac{dR}{dt} &= \gamma((1-f)H + L) \\
\frac{dD}{dt} &= f\gamma H
\end{aligned}$$

where S stands for the susceptible population, R for the recovered population, D for dead, and $H + L$ for the total infections. β is, according to convention the infection rate constant, and γ the recovery rate constant, $0 \le q \le 1$, is the fraction of infections that develop the H manifestation of the disease, and $0 \le f \le 1$ is the fraction of the H disease population that dies from it.

Augmented SIRD model with testing

We go one step further that model the testing for the infection conducted as part of a surveillance program, under the following assumptions:

1. Both uninfected and infected individuals can be tested in a surveillance program, with different probabilities.
2. The amount of surveillance testing capacity is limited to T tests per unit time (day).
3. Uninfected people who are tested, and test negative aren't tested again, unless they show significant symptoms (H) at some point later.
4. Once an individual has tested positive, they are a confirmed case, and any further testing etc. as part of the care program is not counted in this model since such testing will not change the confirmed case numbers, and isn't assumed to come from the surveillance testing capacity.

In total, we have the following states

- S_U, the untested susceptible population,
- H_U, the untested infected highly symptomatic,
- L_U, the untested infected with none or low levels of symptoms,
- R_U, the untested recovered population,
- D_U, the population that died from the disease without being tested
- S_{Tn}, the susceptible population that has been tested, and obviously tested negative,
- H_{Tn}, the highly symptomatic infected population that was earlier tested negative during the susceptible phase (but might be tested in the future during infection)
- L_{Tn}, the low symptom population that was only tested while susceptible, and therefore tested negative at that time,
- H_{Tp}, the high symptom infected population that was tested while in the infected stage, and hence tested positive,
- L_{Tp}, the low symptom infected population that was tested while in the infected stage, and hence tested positive,
- R_{Tn}, the recovered population that was tested in only the susceptible or recovered stages, and hence tested negative,
- R_{Tp}, the recovered population that was tested in the infected population, and hence tested positive,
- D_{Tn}, the deaths due to the epidemic, that were tested negative, and
- D_{Tp}, the deaths due to the epidemic, that were tested positive.

The dynamics from the state transitions is written as

$$
\begin{aligned}
\frac{dS_U}{dt} &= -\beta\frac{S_U\left(L_U+H_U+L_{Tn}+H_{Tn}+L_{Tp}+H_{Tp}\right)}{N} - \min\left(S_U, \frac{S_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right)\\
\frac{dH_U}{dt} &= q\beta\frac{S_U\left(L_U+H_U+L_{Tn}+H_{Tn}+L_{Tp}+H_{Tp}\right)}{N} - \\
&\quad \min\left(H_U, \frac{bH_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right) - \gamma H_U\\
\frac{dL_U}{dt} &= (1-q)\beta\frac{S_U\left(L_U+H_U+L_{Tn}+H_{Tn}+L_{Tp}+H_{Tp}\right)}{N} - \\
&\quad \min\left(L_U, \frac{L_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right) - \gamma L_U\\
\frac{dR_U}{dt} &= \gamma((1-f)H_U+L_U) - \min\left(R_U, \frac{R_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right)\\
\frac{dD_U}{dt} &= \gamma f H_U\\
\frac{dS_{Tn}}{dt} &= -\beta\frac{S_{Tn}\left(L_U+H_U+L_{Tn}+H_{Tn}+L_{Tp}+H_{Tp}\right)}{N} + \min\left(S_U, \frac{S_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right)\\
\frac{dH_{Tn}}{dt} &= q\beta\frac{S_{Tn}\left(L_U+H_U+L_{Tn}+H_{Tn}+L_{Tp}+H_{Tp}\right)}{N} - \\
&\quad \min\left(H_{Tn}, \frac{bH_{Tn} T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right) - \gamma h_{Tn}\\
\frac{dL_{Tn}}{dt} &= (1-q)\beta\frac{S_{Tn}\left(L_U+H_U+L_{Tn}+H_{Tn}+L_{Tp}+H_{Tp}\right)}{N} - \gamma L_{Tn}\\
\frac{dH_{Tp}}{dt} &= \min\left(H_U+H_{Tn}, \frac{b(H_U+H_{Tn})T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right) - \gamma H_{Tp}\\
\frac{dL_{Tp}}{dt} &= \min\left(L_U, \frac{L_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right) - \gamma L_{Tp}\\
\frac{dR_{Tn}}{dt} &= \gamma((1-f)H_{Tn}+L_{Tn}) + \min\left(R_U, \frac{R_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right)\\
\frac{dR_{Tp}}{dt} &= \gamma\left((1-f)H_{Tp}+L_{Tp}\right)\\
\frac{dD_{Tn}}{dt} &= f\gamma H_{Tn}\\
\frac{dD_{Tp}}{dt} &= f\gamma H_{Tp}
\end{aligned}
$$

In addition, the cumulative number of positive and negative tests can be calculated as

$$
\begin{aligned}
\frac{dTp}{dt} &= \min\left(H_U+H_{Tn}, \frac{b(H_U+H_{Tn})T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right) + \\
&\quad \min\left(L_U, \frac{L_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right)\\
\frac{dTn}{dt} &= \min\left(S_U+R_U, \frac{(S_U+R_U)T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right)
\end{aligned}
$$

In case serological tests are done, the cumulative number of positive and negative tests can be calculated as

$$
\begin{aligned}
\frac{dTp^{\text{Sero}}}{dt} &= \min\left(H_U+H_{Tn}, \frac{b(H_U+H_{Tn})T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right) + \\
&\quad \min\left(L_U+R_U, \frac{(L_U+R_U)T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right)\\
\frac{dTn^{\text{Sero}}}{dt} &= \min\left(S_U, \frac{S_U T}{(S_U+L_U+R_U)+b(H_U+H_{Tn})}\right)
\end{aligned}
$$

Schematic, code, and web application to simulate this model is provided in Supplementary Information 2–4.

Estimates of infection and death rates

Using the testing results, the conventional estimate of the infection rate as currently being reported would simply be the fraction of positive test cases found in a time period

$$\widehat{\text{Infection Rate}} = \frac{\Delta T_p}{\Delta T_p + \Delta T_n},$$

and the cumulative death rate estimate would be the calculated from the number of people who died from the pandemic versus those recovered

$$\widehat{\text{Death Rate}} = \frac{D_{Tp}}{D_{Tp}+R_{Tp}}.$$

The case fatality rate, as it is being currently being defined is

$$\text{CFR} = \frac{D_{Tp}}{Tp},$$

which would change to

$$\text{CFR}^{\text{Sero}} = \frac{D_{Tp}}{Tp^{\text{Sero}}}$$

if we use serological testing.

Instead, the true infection rate in the population would simply be the fraction of the population with any kind of infection

$$\text{True Infection Rate} = \frac{L+H}{N},$$

and the true death rate would be simply the total number of people who died of the disease versus the total that died or recovered

$$\text{True Death Rate} = \frac{D}{D+R}.$$

The parameters for the model simulation were based on early data out of Wuhan, as reported by Li et al. [29], shown in Table 1.

Calculating the variance of estimates for unbiased random sampling

Suppose the fraction of population that has contracted SARS-CoV-2 detectable by a serological test (the infection rate) is p. In addition, assume that within those with SARS-CoV-2, a fraction m have died or die within the study time-frame. Therefore, in sampling a random sample of S samples, we expect to find Sp positive cases, and Spm deaths. In terms of the sampled numbers, if we find C positive cases out of a total S sample size and D deaths, the estimates of the infection rate will be $p = C/S$ and the estimate of mortality rate $m = D/C$. These are unbiased estimates, and their conservative, guaranteed confidence interval can be calculated from the Clopper-Pearson interval [30].

Results

An augmented compartment model to estimate epidemic dynamics incorporating testing Bias

We considered an augmentation of the currently prevalent models of epidemic dynamics to explicitly model the potential sampling bias within the tested populations for the specific disease, thus enabling the modeling of reported case numbers and the effects of different testing strategies. Data possibly suggestive of this effect is presented in Fig. 1a. The effect of the sampling bias will tend to decrease with more sampling; in the extreme case of sampling almost the entire population, once the highly symptomatic population is saturated with tests, the rest of the population will start to be sampled at a higher rate, even if the highly symptomatic are being prioritized. Countries with higher tests per positive case tend to show a lower case fatality rate (CFR) (Fig. 1a, SI 1). Indeed, confirmatory rigorous estimates of the sampling bias will be possible from simultaneous unbiased random sampling studies in parallel with the conventional method of mainly testing suspected cases and their contacts. A similar effect is observed longitudinally for the calculated CFRs and testing in the US (Fig. 1b). Anti-correlation of CFR across countries with higher sampling in a given period, and with progressively increased sampling at a given location over time underline the existence of sampling bias for patients with high disease manifestation. In this paper, we asked how such sampling bias might be affecting the prediction of epidemic dynamics. We therefore explicitly incorporated the sampling bias by stratifying infected population into the highly symptomatic and the low symptom/asymptomatic, in the commonly used SIRD model for epidemic dynamics, with an objective to demonstrate the effect of sample bias in the reported case numbers and any subsequent direct estimates of infection and fatality rate dynamics using simulations.

The most common model used to study epidemic dynamics is the SIRD compartment model, from which many derivatives have been designed. We decided to choose the simplest SIRD model to test whether directly incorporating the testing bias could have an effect in the estimate of patients belonging to a given compartment. The basic SIRD compartmental model stratifies the population in 4 compartments: Susceptible (S), Infected (I), Recovered (R), and Dead (D). Movements of subpopulations from one compartment to the other are described by ordinary differential equations (ODEs). The parameters for each of these ODEs are the rate constants, β, describing the rate of infection, and γ, describing the rate of death (Fig. 1c).

In various countries, the initial tests for viral presence have been biased based on the severity of the disease manifestation, or weighted towards symptomatic patients. However, in certain situations, these biases could be present in other directions too, wherein patients with

Table 1 Parameters used in the model

Parameter	Meaning	Value	Source	Notes on calculation from data source
β	Transmission rate	$1.12 \times \frac{3.47}{3.47+3.69}$ days^{-1} = 0.543 days^{-1}	Li et al. [29]	Calculalated by weighting the transmission rate by the latency and infectious periods
γ	Disease recovery or resolution rate	$\frac{1}{(3.69+3.47)\ \text{days}} = \frac{1}{7.16\ \text{days}}$	Li et al. [29]	Summing up the latency and infectious period
q	Fraction of infected individuals who are symptomatic (or with more severe disease)	0.15	Li et al. [29]	about 86% were undocumented infected
f	Fatality rate for individuals with severe/ symptomatic disease	0.02	–	Nominal value for illustrative purposes

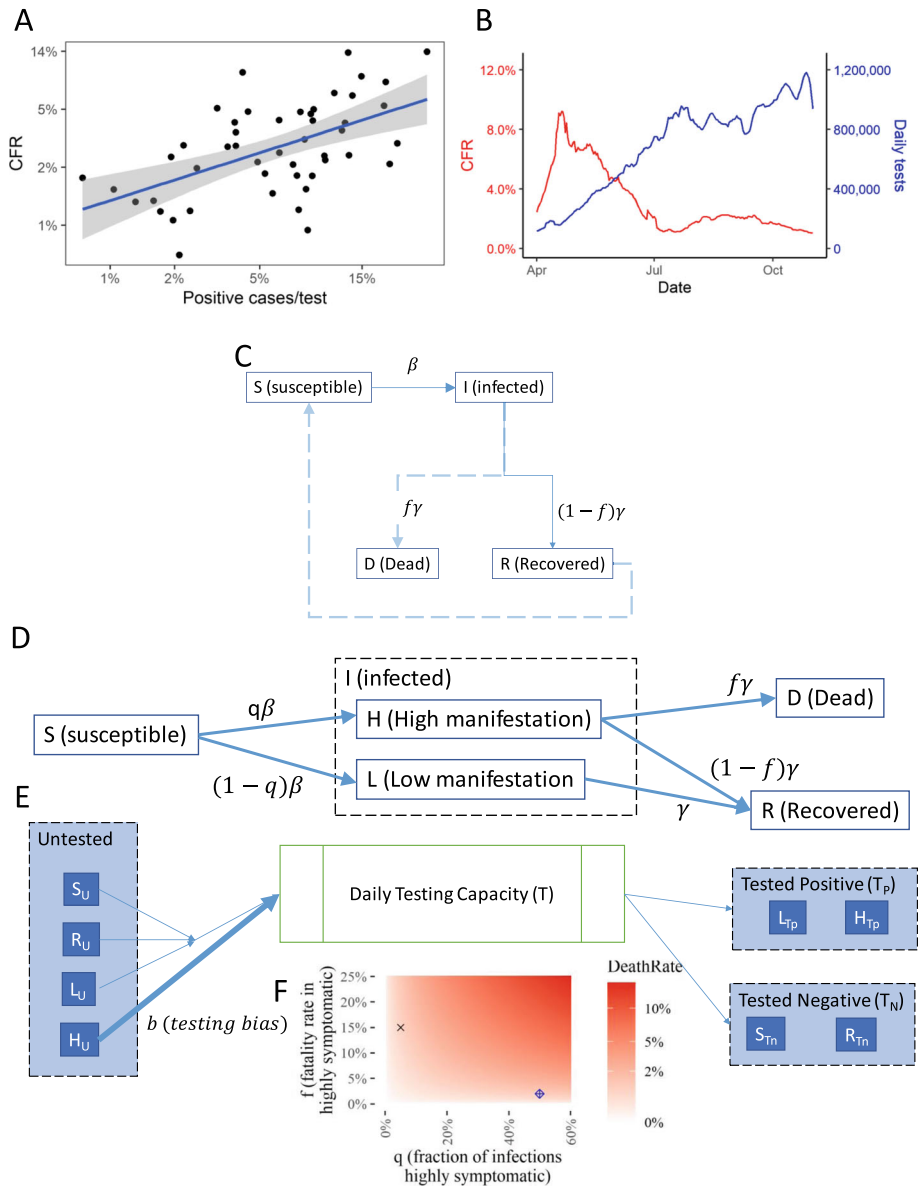

Fig. 1 An Augmented Compartment Model to Predict Epidemic Dynamics with Testing Bias. **a** Regression of Case fatality rate, CFR (calculated as percentage of death in positively identified cases per country) against the percentage of positive cases identified among all tested per country show a linear regression; Each dot corresponds to a different country; Data obtained from ourworldindata.com for April 18, 2020 (SI 1); Blue line shows the fitted regression curve; Shaded area show the 95% confidence interval; $R^2 = 0.3567$, *p*-value = 5.9e-6. **b** Longitudinal data of the calculated CFR vs. the total number of individuals tested daily in the US, with the tests performed, deaths reported, and new cases confirmed smoothed using the 7-day averages. **c** The basic SIRD compartmental model commonly used to model epidemic dynamics. Ordinary differential equations describe the movement of the population through the different compartments representing the susceptible, infected, recovered, and dead stages. The parameters are the rate constants for each term representing the transitions in the differential eq. **d** The augmented SIRD model by stratification of the infected population into H and L referring to high, and low manifestation of disease symptoms respectively; The factor q is the fraction of infected within H; We assume that high manifestation of disease leads to death in a fraction f of the individuals. **e** A simplified representation of the model of the testing policy; T tests are available per unit time; Untested alive individuals (U) are randomly selected in proportion to their numbers, but patients in H are selected with an increased bias b. Further compartments arising due to testing and movement at different stages are omitted here for clarity; Detailed equations in the Methods section. (**e**) Fraction f and q determine the true death rate; Two values with similar death rates chosen for simulations are marked

more likelihood of death are under-sampled. We therefore decided to introduce testing bias by stratifying the infected population based on the severity of disease manifestation. Specifically, we further stratified the infected compartment (I) into two other sub-compartments, H (high) and L (low) referring to the

high or low symptomatic manifestation of the disease respectively. Although the transition from the S (susceptible) compartment to the infected (I) is driven by rate constant β, the factor q describes the fraction of the infected subpopulation with a high symptomatic manifestation of the disease. We assumed that the fraction within L (low manifestation) die in miniscule rates, and nearly all deaths occur from the H fraction. This added sub-compartmentalization is a simple addition to the model, but we believe that if well-defined stratification could be measurably identified within the infected (I) compartment, more sub-compartments should be added. These fractions could include people with known comorbidities with a higher chance of fatality, or those with measurably high severity of disease manifestation.

We then superimposed upon our augmented compartment model a testing policy (Fig. 1d). We assumed that T tests are available per unit time (kept constant for simulations below, but which could itself be a time varying function based on the availability of testing capabilities over time). The untested, and alive individuals are assumed to be randomly selected for testing in proportion to their numbers, but those with high disease manifestation (H) are selected with an increased bias b. In addition, patients who were tested as being negative for viral load at a previous time point, but presenting severe symptoms at the present time would also be selected with an increased bias b. The biased testing policy was implemented by splitting the compartments in our augmented SIRD model for the untested and the tested fractions (Fig. 1e). True death rate would depend upon the factors q (fraction with high disease manifestation), and f (fraction dying within the H compartment) (Fig. 1f). Details for the ODEs describing the transition through these compartments are provided in the Methods section.

Testing Bias strongly affects the direct estimation of infection rate

We simulated our augmented compartmental model with testing bias and calculated the infection rate dynamics based on an active viral test (based on measurement of viral sequences), as well as based on a serological test (measuring if antibodies against the virus have been created, or more accurately directly testing for a T cell mounted response). In our augmented model, the estimated infection rate is calculated as a ratio of those tested positive, and all tested population within a given time frame. Here, the testing bias is reflected within the sampling of stratified populations, H and L in the infected (I) compartment.

We found that testing bias had a profound effect on the naïve estimation of infection rate based on active viral test, the most commonly employed tests (Fig. 2a). In contrast, estimation of the contraction rate was much less affected by the bias, largely because the immunologically recovered population as a fraction of the total population increases as time progresses (Fig. 2b). Our simulation provides a strong argument in favor of serological testing beyond the obvious argument of their capability to correctly assign the compartment of recovered (R) fraction to the population which contracts the disease but tests negative. That the testing bias could be substantially mitigated in the estimation of contraction rate by serological test is a strong argument in favor of serological testing, although these tests are unlikely to be available in the initial spread of a new epidemic. Nevertheless, our augmented model will allow estimation of the effect of biased sampling itself in predicting disease dynamics, and underlines the importance of unbiased sampling to predict estimates reflecting reality.

Testing Bias influence true fatality rates and case fatality rates in time-dependent manner

Since the initial report of case fatality rate of 3.8% from Wuhan China by WHO, there has been a substantial variance in the listed death rate among nations. Case fatality rate is calculated as the ratio of number of deaths measured and number of positive cases [2]. Since death is a lag indicator, case fatality rate would asymptotically reach the more conservative death rate, which is the ratio of deaths and a sum of those who died or recovered. We, therefore, considered the latter death rate and tested the effect of sampling bias upon its estimation. We found that sampling bias can linearly affect the estimation of death rate (Fig. 2c). If the fraction of population with high manifestation of disease is high (and therefore the inherent bias of sampling is low), then trivially the effect of bias is somewhat mitigated (Fig. 2d). Our data underlines the importance of incorporating bias in testing itself as a key parameter to model epidemic dynamics, and characterizes the effect of bias on key predicted metrics, including death rate, are substantially affected by these biases.

We then tested how testing bias would affect the case fatality rate (ratio of dead to the number of positive cases). We found that sampling bias indeed resulted in large effects in CFR estimates, but crucially, these effects reduce as the infection reaches its peak, and then amplify as the infections subside within the population (Fig. 2e). When compared to the true death rate, CFR initially underestimates the death rate, and then overestimates the rate in a bias dependent manner. CFR, as is calculated here, has two opposing biases inherent in it. In the initial part of the pandemic, and for testing biases below a threshold, it underestimates the true death rate because the growth of the infections happens earlier than the growth in the number of death. In other words, while the pandemic is growing, the number of deaths always lag, and the number of infections at a particular

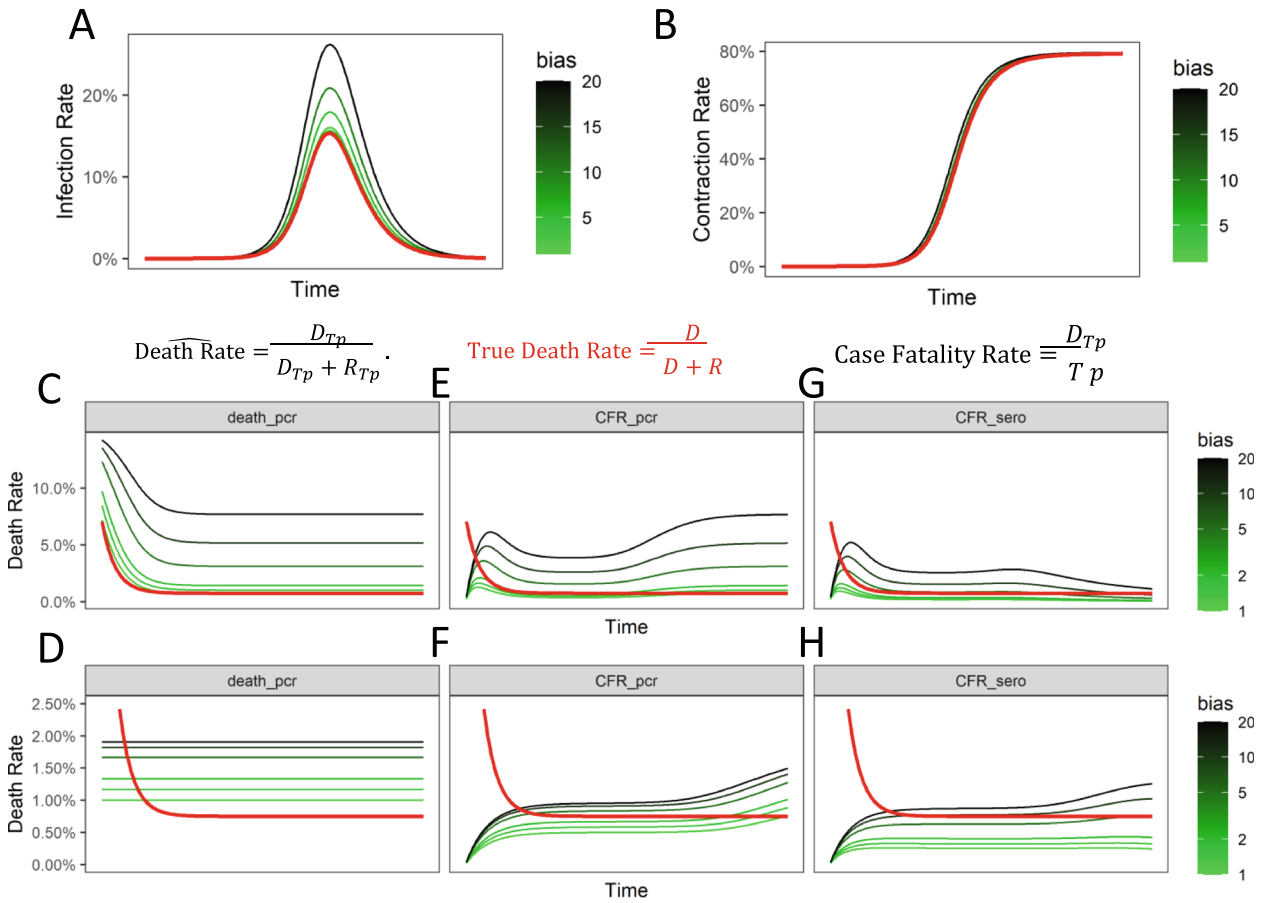

Fig. 2 Influence of testing (sampling) bias on estimates of disease spread and death rates for symptomatic patients is high for estimation of infection rate measured by active-viral tests, but mitigated in serological tests for acquired immunity. **a** The infection rates (fraction of population with an active infection) estimated from PCR (or sequencing) based strategy to measure active viral load; Red curve shows the true infection rate; Estimation of infection rates with different biases for H-compartment patients (those with higher disease manifestation) shown in green-black lines; Also shown are fractions calculated in eqs. **b** Estimation of contraction rates (fraction of population that has contracted the virus at a previous time, are either infected or recovered) from a serological test; Red line is true contraction rate; Green-black lines are estimated contraction rates with different sampling biases for patients with higher disease manifestation. **c-d** Estimates of death rate (percentage of fatalities within confirmed recovered and dead patients) from a biased sampling strategy for patients with high disease manifestation at different fractions (**c**) 15% of the infected population with 2% death rate obtained from previously published studies, and (**d**) 50% of the infected population with .5% death rate; Red lines refer to true death rate and green-black lines refer to biased estimates of death rates. **e-f** Estimates of case fatality rate (percentage of confirmed fatalities within positively tested population) from a biased sampling strategy as in C, and D for different fractions of population in H and L manifestation, as in **c,** and **d**. **g-h** Estimates of case fatality rate measured with serological testing from a biased sampling strategy for different fraction of populations in H and L manifestation, as in **c**, and **d**

time are some multiplicative factor larger than the corresponding infections that existed when the currently dead were infected. This leads to estimation of an overly optimistic CFR. On the other hand, for later and waning stages of the epidemic, and the testing biases being greater than a threshold, the CFR leads to an overly pessimistic number compared to the real death rate. This is due to a greater prevalence of the severely ill patients counted among the cases. Indeed, if the fraction of population with a high disease manifestation (H) are changed and correspondingly the death rate adjusted to keep the true death rate the same, then CFR can underestimate the true death rate for a longer duration of the pandemic (Fig. 2f). The direction of the bias in the CFR depends on the pandemic kinetics (β, γ) and the testing bias.

However, if CFR were to be measured using serological tests, thereby counting the recovered population as being previously infected, the effect of bias on estimation of fatality rate is mitigated (Fig. 2g-h). Crucially, our calculations argue that for all metrics of naïve estimation of fatality rates, contribution of testing bias is substantial, and attempts made to measure the bias itself, and be accounted for.

Randomized unbiased serological sampling of widely infected population is necessary to determine true fatality rates

Our augmented model demonstrated that sampling bias could play a significant role in the direct estimates of both the infection and the fatality rates, and may partially explain the large variance across the death rates reported across countries, as well as in epidemic prediction models. A random sampling of a population with a large infection load could be utilized to estimate the true infection, recovery, and fatality rates. Here, we provide calculations for the sample sizes required to gain an accurate estimate of the community infection rates and the infection fatality rates. In order to minimize the variance of the infection death rate, the random testing is suggested to be conducted among a population wherein the infection is understood to have spread widely. A random selection of individuals with an unbiased identifier (e.g. tax ID) will provide estimates without systematic biases; therefore, the appropriate measure of accuracy need only be concerned with the variance of the estimates. In the following, we have chosen to frame this in terms of mostly confidence intervals and hypothesis testing.

An initial calculation expectedly suggested that with low infection rates, attaining a 5% error of estimation for the infection rate would require a moderate sample size. In contrast, if the real infection rate is higher (closer to 50%), expectedly, a smaller sample set will be sufficient for an accurate estimate of infection rate (Fig. 3). Therefore, in the present scenario, an example of an ideal location where such tests could be performed with a limited number of sample size (approximately 10,000) is Germany or India, where the deaths have rapidly climbed up in July. Crucially, in India, a randomized

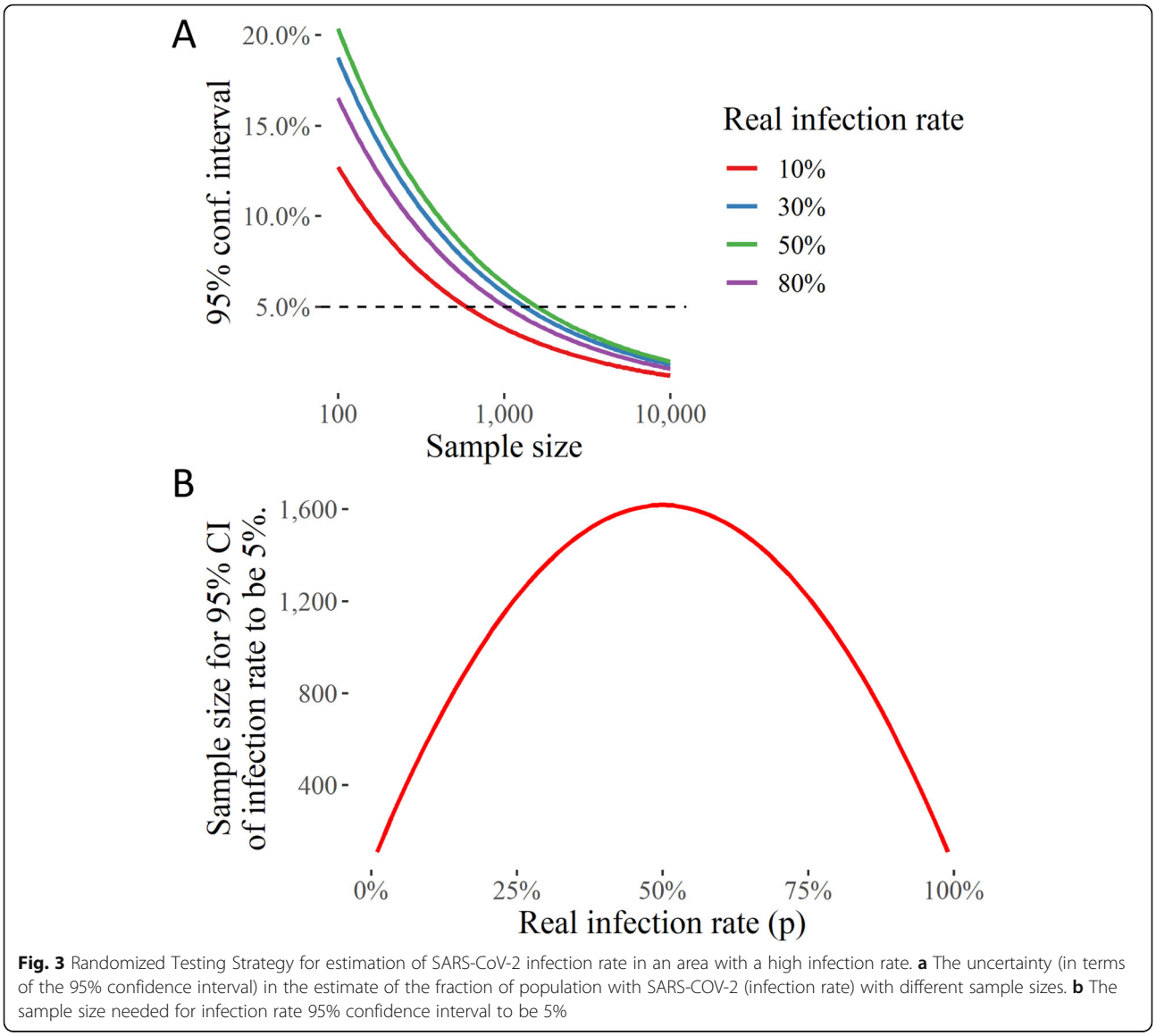

Fig. 3 Randomized Testing Strategy for estimation of SARS-CoV-2 infection rate in an area with a high infection rate. **a** The uncertainty (in terms of the 95% confidence interval) in the estimate of the fraction of population with SARS-COV-2 (infection rate) with different sample sizes. **b** The sample size needed for infection rate 95% confidence interval to be 5%

testing indicates that a quarter of Delhi's population may have already contracted the virus, and a nearly similar percentage in a much more crowded Mumbai [31, 32].

Estimating the mortality or infection fatality rate requires another probability to be multiplied to the estimate of infection rate within a sample population. A calculation of the sample size required for a 95% confidence interval indicates that even for a potentially highly infected population, like in NYC, it may require a very large sample size to accurately determine the true fatality rate (Fig. 4a). This

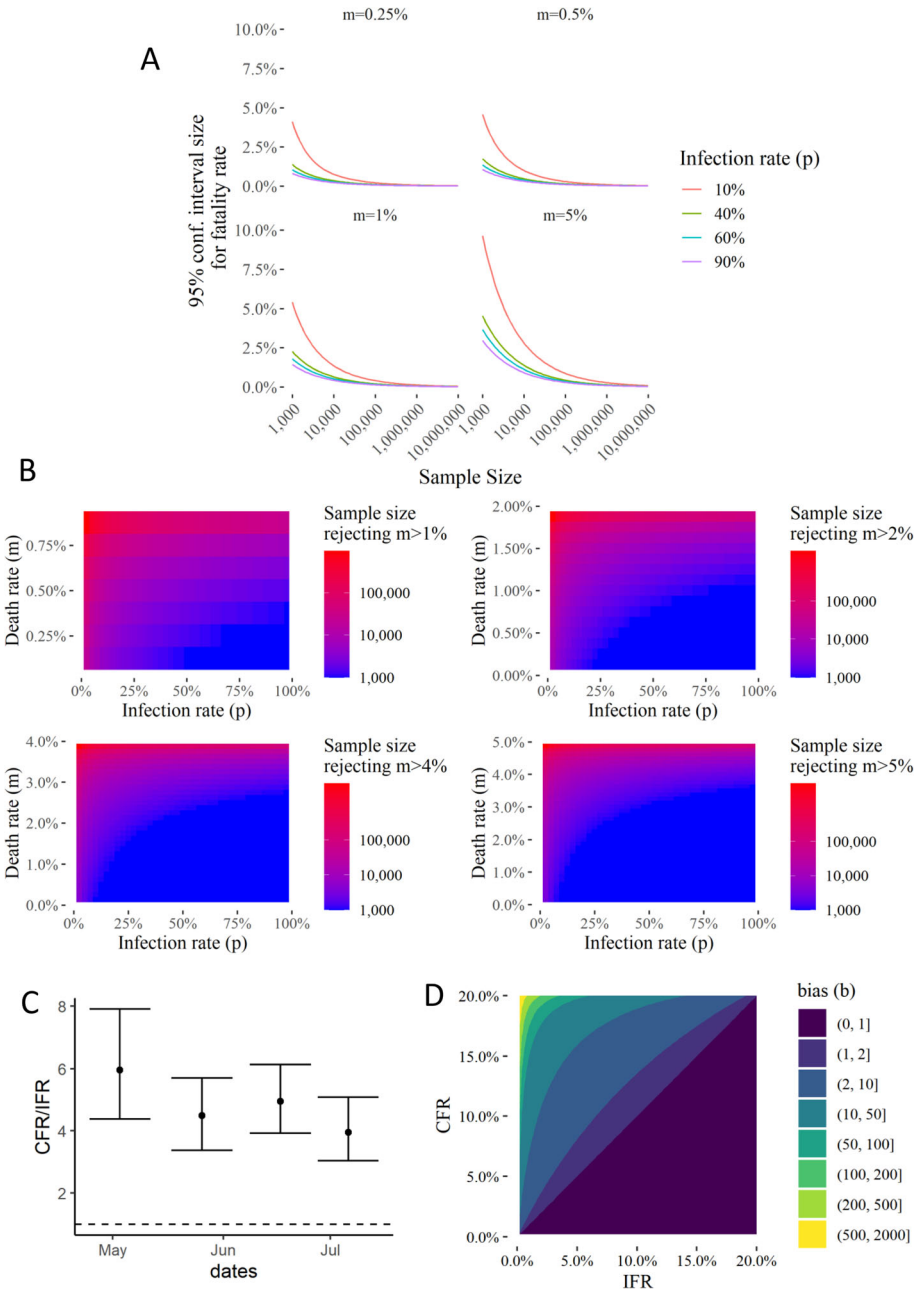

Fig. 4 Randomized Testing Strategy and comparison with biased sampling for estimation of fatalities, and for estimate of death rate below a given percentage in areas with different infection and fatality rate. **a** 95% Confidence Interval size for the death rate given a sample size (S, x-axis), infection rate (p, line colors), and the real death rate (m, subplot panel). **b** Sample sizes needed to reject hypotheses that death rate > than 1, 2, 4, or 5% of the infected population. **c** Ratio of the CFR calculated from all tests (biased sampling), to the Infection Fatality Rate (IFR) calculated from CDC's seroprevalence study in CT. Serosurvey IFR is calculated by estimating the fraction of the state population having contracted the virus. **d** Comparison of the CFRs calculated from a hypothetical unbiased random sampling and biased sampling as a function of the sampling bias. Testing bias could be inferred from unbiased random sampling using this analysis

may be one reason why countries have resorted to large sampling to obtain data for true fatality rate. However, biased and non-random sampling renders these data difficult to interpret to estimate the fatality rates.

We therefore propose to instead test the hypotheses that the true fatality rate is higher than a given value, which would be rejected if the upper limit of the 95% confidence interval is lower than the said value. Calculating these sample sizes with the statistically significant 95% confidence interval, we found that a relatively much smaller sample size would be sufficient to estimate if the true fatality rate is below or higher than a given percentage (Fig. 4b). Our calculations indicate that for a sample with 50% infection rate, a sample size of 1000 may be sufficient to identify if fatalities are much lower than 1%, while for a sample with 25% infection rate, it may be below 10,000 — a logistically achievable size to determine a crucial parameter.

Continuous sampling of a selected cohort can provide useful dynamics on acquirement of immunity

The availability of a serological test, if applied using a random and unbiased sampling strategy could allow the identification of a key subset of people who have developed immunity, but do not carry the infectious disease burden. However, it is usually not possible to have antibody tests available at the onset of a disease, and a rapidly spreading pandemic may make it difficult to gear policies based on an accurate assessment of the development of herd immunity. In contrast, the recent development of genomic amplification or sequencing technologies has made it possible to prepare rapidly deployable tests to assess active infectious loads. We therefore propose to use a continuous sampling of a representative unbiased cohort on a weekly basis to determine the initial onset of infection, the rate of its spread, development of immunity, and eventually the ensuing aftermath of the infection. Indeed, as we showed, for very small infection rate, a larger sample may be required. However, this concern is easily addressable by pooled sequencing (NGS), which can be used to determine rare onset, mutagenesis, and characterization of infections [33–35], and if sufficient signal for infection is found, then the continuous sampling be used for that cohort. The dynamics of readout (of active viral load) in a fixed sample set will allow an accurate estimation of the development of immunity and its dynamics in a given population.

Inferring bias from a comparison of biased and unbiased sampliong

Since the effect of the biased sampling should be apparent in the overestimation of case fatality rates, we compared the CFRs calculated from the total tests with 4 seroprevalence surveys [36] conducted by the CDC in Connecticut. While these seropravalence surveys were not truly unbiased, they tested blood samples collected by diagnostic labs for reasons unrelated to COVID19. CFRs calculated from total tests were always higher than the CFRs calculated from the seroprevalence surveys (Fig. 4c). Finally, we calculated the CFRs that would be seen by hypothetical unbiased sampling vs those from biased sampling, as we vary the fraction of high and low symptomatic populations, while keeping the death rate of the high symptom population at 20%. Since the followup of unbiased tested individuals will allow for estimates of both these death rates and fractions of the infected populations, we can infer the sampling bias, as seen in Fig. 4d. Thus, truly unbiased random sampling with symptoms and outcome tracking of patients can not only provide accurate estimates of death rates and disease dynamics, comparing the results to conventional testing can also provide estimates of the sampling bias.

Discussion

The wide, and constantly updated, estimates of key metrics of the disease, including fatality and recovery rates associated with SARS-CoV-2 raise important questions about the quality of our public health scientific inquiry. Additionally, the effects of the public health and economic policies adopted around the world on the socio-economically and politically vulnerable sections of the population has so far received insufficient attention. This is the most severe global health crisis to have inflicted humanity within this generation, although its true impact has still not been understood completely. Indeed, even after months of its spread, there is a large variation in the estimate of infection, recovery, and fatality rates. Crucially, an accurate estimate of the true dynamics and infection, recovery, and fatality rates is necessary for the scientific inquiry into the disease from a systems perspective, operations planning and to advocate for an apt public health policy. We show that direct estimates of these parameters lack in this respect methodologically. Testing of the general population has been incidentally biased towards symptomatic patients, since it is driven by the desire to identify, contact trace, care for, or safely isolate vulnerable populations rather than to estimate accurate metrics. This has resulted in the biased sampling that is not well suited for modeling of the epidemic and calculation of key metrics. We provide calculations to ascertain the estimate of bias from independent unbiased surveys conducted explicitly for its own purpose, to provide useful “corrections” to the inaccurate epidemic predictions from the incidentally biased surveys.

Conclusions

In this work, we attempted to systematically characterize the crucial issue of testing bias by incorporating the testing bias within the compartmental model of epidemic dynamics. Our model also includes tests for active viral load, as well as for those who have developed immunity, along with the sampling bias in testing. We believe that our proposed augmentation not only provides a systematic basis to ascertain the effect of testing policies in estimates of epidemic dynamics, but also demonstrates that biased sampling may substantially influence epidemic projections if metrics are naïvely calculated only reported case numbers.

Another problem that we demonstrate with the directly calculated case fatality rates is that the counts of deaths lag the infection rates, and therefore while the pandemic is growing, the case fatality rates underestimate the infection fatality rates. While the bias introduced due to this lag is in the opposite direction to that introduced by the sampling bias, the result only makes the situation worse. In terms of statistical theory, the case fatality rate is a large variance, biased estimate with an unknown direction of bias. In the popular press discussions, case fatality rate estimates calculated for countries with proportionately very extensive testing such as Iceland have been optimistically cited as the true infection fatality rates [37, 38]. However, as some sources of systematic errors are mitigated in extensive testing, the effect of the lagging death counts will proportionately become more important. Therefore, until we arrive towards the end of the pandemic, these optimistic case fatality rates may be more optimistic than the reality.

Although much data is collected on the number of cases, the ensuing deaths, and those that have recovered, the naïve interpretation of fatality and infection rates from non-uniform sampling across countries may be fraught with substantial inherent problems. Therefore, we recommend a limited, unbiased, random uniform sampling of population to test hypotheses of fatality rates. We also propose a method to continually monitor a static sample set to estimate the onset, and dynamics of disease spread, acquired immunity, and ensuing morbidity and fatalities associated with an infectious spread. As a recent example, a large number of deaths in New York City could be explained either by (i) a high fatality rate in a small population contracting the virus, or (ii) a rapid spread of the virus which has resulted in large number of people to develop immunity with a smaller percentage succumbing to the viral infection. In order to distinguish between the two widely varying scenarios, the most direct method with the least amount of statistical assumptions, would be to serologically test a limited, random sample of individuals. This should occur in addition to any surveillance methods currently being employed that are independently needed for targeted medical care and public health interventions. Disease prediction from the incidentally biased sampling can then be corrected for by ascertaining the extent of bias from fatality rates independently derived from unbiased sampling.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-020-01196-4.

Additional file 1 SI 1. Table of CFRs and tests per positive case for different countries.

Additional file 2 SI 2. Transition diagram of the full compartmental model incorporating SIRD dynamics and biased testing.

Additional file 3 SI 3. R Code to simulate the augmented SIRD model with biased sampling.

Additional file 4 SI 4. Webapp to simulate the augmented SIRD model with biased sampling (https://yasir.shinyapps.io/biased__sampling_SIRD/).

Abbreviations

SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2; COVID-19: Coronavirus Disease 2019

Acknowledgements

Not applicable.

Authors' contributions

K and YS conceived the idea, YS created the statistical groundwork and generated the figures, JA provided the medical rationale and helped in writing the manuscript with all other authors. All authors have read and approved the manuscript.

Funding

Funding was provided by the UConn Health Startup Funds. The funding body (University of Connecticut) had no role in the design of the study, collection, analysis, interpretation of data or in writing the manuscript.

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Biomedical Engineering, University of Connecticut Health, Farmington, CT, USA. [2]Center for Cancer Systems Biology @ Yale, West Haven, CT, USA. [3]Department of Medicine, University of California, San Francisco, CA, USA.

Received: 28 July 2020 Accepted: 18 December 2020
Published online: 07 January 2021

References

1. WHO: Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) 2020.
2. Spychalski P, Błażyńska-Spychalska A, Kobiela J. Estimating case fatality rates of COVID-19. Lancet Infect Dis. 2020;20(7):774-5.
3. Kim DD, Goel A. Estimating case fatality rates of COVID-19. Lancet Infect Dis. 2020;20(7):773-4.
4. Lipsitch M. Estimating case fatality rates of COVID-19. Lancet Infect Dis. 2020;20(7):775.
5. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. Lancet Infect Dis. 2020;20(7):773.

6. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet. 2020;395(10225):689–97.
7. Read JM, Bridgen JRE, Cummings DAT, Ho A, Jewell CP. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. medRxiv. 2020. https://doi.org/10.1101/2020.01.23.20018549.
8. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med. 2020;382(13):1199–207.
9. Should We have Locked the World Down for the COVID-19?. https://medium.com/@kshitizkz/should-we-have-locked-the-world-down-for-the-covid-19-e0dc5191034c. Accessed 29 Dec 2020.
10. Kshitiz: Should We Have Locked The World Down? https://swarajyamag.com/ideas/should-we-have-locked-the-world-down. 2020.
11. Wolfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Muller MA, Niemeyer D, Jones TC, Vollmar P, Rothe C, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020;581:465–9.
12. To KK, Tsang OT, Leung WS, Tam AR, Wu TC, Lung DC, Yip CC, Cai JP, Chan JM, Chik TS, et al. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. Lancet Infect Dis. 2020;20(5):565–74.
13. Sood N, Simon P, Ebner P, Eichner D, Reynolds J, Bendavid E, Bhattacharya J. Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020. JAMA. 2020;323(23):2425-7.
14. Leuzinger, K, Gosert, R, Søgaard, KK, et al. Epidemiology and precision of SARS-CoV-2 detection following lockdown and relaxation measures. J Med Virol. 2020;1–11. https://doi.org/10.1002/jmv.26731. Accessed 29 Dec 2020.
15. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, Rawlings SA, Sutherland A, Premkumar L, Jadi RS, et al. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell. 2020;181(7):1489–501 e1415.
16. Mathew D, Giles JR, Baxter AE, Oldridge DA, Greenplate AR, Wu JE, Alanio C, Kuri-Cervantes L, Pampena MB, D'Andrea K, et al. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science. 2020;369(6508):eabc8511. https://doi.org/10.1126/science.abc8511.
17. Braun J, Loyal L, Frentsch M, Wendisch D, Georg P, Kurth F, Hippenstiel S, Dingeldey M, Kruse B, Fauchere F, et al. Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. medRxiv. 2020. https://doi.org/10.1101/2020.04.17.20061440.
18. Nickel CH, Rueegg M, Pargger H, Bingisser R. Age, comorbidity, frailty status: effects on disposition and resource allocation during the COVID-19 pandemic. Swiss Med Wkly. 2020;150:w20269. https://doi.org/10.4414/smw.2020.20269. Accessed 29 Dec 2020.
19. Russell TW, Hellewell J, Jarvis CI, van Zandvoort K, Abbott S, Ratnayake R, Cmmid Covid-Working G, Flasche S, Eggo RM, Edmunds WJ *et al*: Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the diamond princess cruise ship, February 2020. Euro Surveill 2020, 25(12).
20. Lourenco J, Paton R, Ghafari M, Kraemer M, Thompson C, Simmonds P, Klenerman P, Gupta S: Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. medRxiv. 2020. https://doi.org/10.1101/2020.03.24.20042291. Accessed 29 Dec 2020.
21. Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, Bhatia S, Boonyasiri A, Cucunubá Z, Cuomo-Dannenburg G *et al*: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. . Imperial College COVID-19 Response Team 2020.
22. Sanchez-Caballero S, Selles MA, Peydro MA, Perez-Bernabeu E. An Efficient COVID-19 Prediction Model Validated with the Cases of China, Italy and Spain: Total or Partial Lockdowns? J Clin Med. 2020;9(5):1547.
23. Garcia-Basteiro AL, Chaccour C, Guinovart C, Llupia A, Brew J, Trilla A, Plasencia A. Monitoring the COVID-19 epidemic in the context of widespread local transmission. Lancet Respir Med. 2020;8(5):440–2.
24. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun. 2020;11:5749. https://doi.org/10.1038/s41467-020-19478-2.
25. Nickel CH, Bingisser R. Mimics and chameleons of COVID-19. Swiss Med Wkly. 2020;150:w20231.
26. Zhao Q, Ju N, Bacallado S: BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic. arXiv preprint arXiv:200407743 2020.
27. Verity R, Okell LC, Dorigatti I, Winskill P, Withtaker C, Imai N, Cuomo-Dannenburg G, Thompson H, Walker P, Fu H, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis. 20(6):669-77. https://doi.org/10.1016/S1473-3099(20)30243-7.
28. Brunner J, Chia N. Confidence in the dynamic spread of epidemics under biased sampling conditions. PeerJ. 2020;8:e9758.
29. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020;368(6490):489–93.
30. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404–13.
31. Thacker T. ICMR intensifies random testing to check for covid-19. In: The Economic Times. 03.16; 2020.
32. Marpakwar C: ICMR approves four more testing centres in Mumbai. In: Mumbai Mirror. Mumbai: Times of India Group; 2020.
33. Peddu V, Shean RC, Xie H, Shrestha L, Perchetti GA, Minot SS, Roychoudhury P, Huang ML, Nalla A, Reddy SB, et al. Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral Superinfection and colonization. Clin Chem. 2020;66(7):966–72.
34. Sibley CD, Peirano G, Church DL. Molecular methods for pathogen and microbial community detection and characterization: current and potential application in diagnostic microbiology. Infect Genet Evol. 2012;12(3):505–21.
35. Skums P, Artyomenko A, Glebova O, Ramachandran S, Mandoiu I, Campo DS, Dimitrova Z, Zelikovsky A, Khudyakov Y. Computational framework for next-generation sequencing of heterogeneous viral populations using combinatorial pooling. Bioinformatics. 2015;31(5):682–90.
36. Havers FP, Reed C, Lim T, Montgomery JM, Klena JD, Hall AJ, Fry AM, Cannon DL, Chiang CF, Gibbons A, et al. Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. JAMA Intern Med. 2020. https://www.the-sun.com/news/671990/iceland-coronavirus-testing-reveals-less-deadly-half-population-asymptomatic/. Published April 12, 2020.
37. Sullum J. What We Should Have Learned From Iceland's Response to COVID-19. In: reasoncom. 4.3; 2020.
38. Lock S. Coronavirus may be LESS deadly than we thought as Iceland testing reveals huge numbers had disease without realising. In: The US Sun. 4.29; 2020.

Publisher's Note