Evaluation of the number of COVID-19 undiagnosed infected using source of infection measurements

Akiva B Melka[1], Yoram Louzoun[1,2,]*

[1]Department of Mathematics, Bar-Ilan University, Ramat Gan 52900, Israel
[2]Gonda Brain Research Center, Bar-Ilan University, Ramat Gan 52900, Israel
[*]Corresponding author: louzouy@math.biu.ac.il

Abstract

Multiple studies have been conducted to predict the impact and duration of the current COVID-19 epidemics. Most of those studies rely on parameter calibration using the published number of confirmed cases. Unfortunately, this number is usually incomplete and biased due to the lack of testing capacities, and varying testing protocols. An essential requirement for better monitoring is the evaluation of the number of undiagnosed infected individuals. This number is crucial for the determination of transmission prevention strategies and it provides statistics on the epidemic dynamics. To estimate the number of undiagnosed infected individuals, we studied the relation between the fraction of diagnosed infected out of all infected, and the fraction of infected with known contaminator out of all diagnosed infected. We simulated multiple models currently used to study the COVID-19 pandemic and computed the relation between these two fractions in all those models. Across most models currently used and for most realistic model parameters, the relation between the two fractions is consistently linear and model independent. This relation can be used to estimate the number of undiagnosed infected, with no explicit epidemiological model. We apply this method to measure the number of undiagnosed infected in Israel. Since the fraction of confirmed cases with a known source can be obtained from epidemiological investigations in any country, one can estimate the total number of infected individuals in the same country.

Introduction

Since the initial spread of COVID-19, the number of infected individuals has been used as one of the main tools to measure the spread of the virus.[1,2] However, it is now clear from measures in China and around the world that there is a substantial number of undiagnosed infected.[3—5] To estimate the total number of infected, it is crucial to determine the Confirmed Cases Fraction (*CCF*), which is defined here as the fraction of confirmed (diagnosed) infected out of all infected (both diagnosed and undiagnosed). A precise estimate of *CCF* is essential for the assessment of the current situation in any given country and the establishment of protective measures. In different countries, *CCF* values depend on sampling protocols and frequencies. Except for Iceland that has now tested its entire population and identified all infected individuals, there are obvious disparities between countries in their testing capacities and protocols[6,7]. Recently, serology-based estimates of the total exposed number emerge.[28,29] However, it is not clear that these exposed are infective.

In the absence of a vaccine or efficient treatment, the control of social contacts through large-scale social distancing measures appears to be the most effective means of mitigation in the current COVID-19 pandemic. Modeling has emerged as an important tool to gauge the potential for widespread contagion, to cope with associated uncertainty, and inform its mitigation, and has increased the interest in epidemiological models.[8—13] The predictions of those models are often based on observations-driven parameter estimates. However, one of the main observations for this pandemic - the reported number of carriers is heavily influenced by sampling biases. To correlate the number of reported diagnosed and the total number of infected, one must estimate *CCF*.

Multiple models have been proposed to evaluate *CCF* using, for instance, the number of deceased patients,[14,15] but in all those studies, the results depend on the models used or on estimates of country-specific parameters, such as the age dependence or the Infection fatality rate (*IFR*). We here propose a novel approach to evaluate *CCF* based on the fraction of diagnosed infected with a known infection source (Known Source Fraction *KSF*). We show that this method is not sensitive to the details of the model used.

Two main types of predictive models were proposed for the current COVID-19 epidemic: macroscopic models, using aggregated data at the population scale and microscopic models, incorporating distributed information at the individual level.[16—18] Macroscopic models use stochastic processes or ODEs to predict the evolution of the outbreak on a global scale.[11,16] The simplest and most common model is the SIR model,[19] where the population is divided into three categories: Susceptible, Infected, and Removed (figure 1 upper scheme). In this model, propagation of the virus depends on the infection rate or the number of contacts between susceptible and infected individuals, and the detection rate that characterizes the time that infected individuals remain contagious. The Removed category can include individuals that survived the virus and are now immune, or deceased patients (SIRD). If stringent confinement is applied, this category can also simply be all diagnosed individuals since they are now removed from the system and can no longer contaminate other individuals.

A difference between COVID-19 and previous respiratory diseases is the relatively long incubation time. Therefore, a fourth category is often added: Exposed individuals (SEIR model – figure 1 second scheme) that carry the virus but still do not contaminate others. A lag of a few days has indeed to be considered when observing infection patterns.[21,22] More sophisticated versions of SEIR also incorporate migration to assess the efficiency of intercity restrictions.[23] In light of the latest development, other categories can be added to the system such as Asymptomatic individuals. Finally, models were refined with a time-dependent infection rate,[17] age-dependent infection matrices, or even quarantine.[24,25] The practical conclusions drawn from such models rely on predetermined assumptions on the *CCF* over time. We here propose a straightforward method to estimate the time-dependent *CCF* and show that it is not sensitive to the details of the model used.

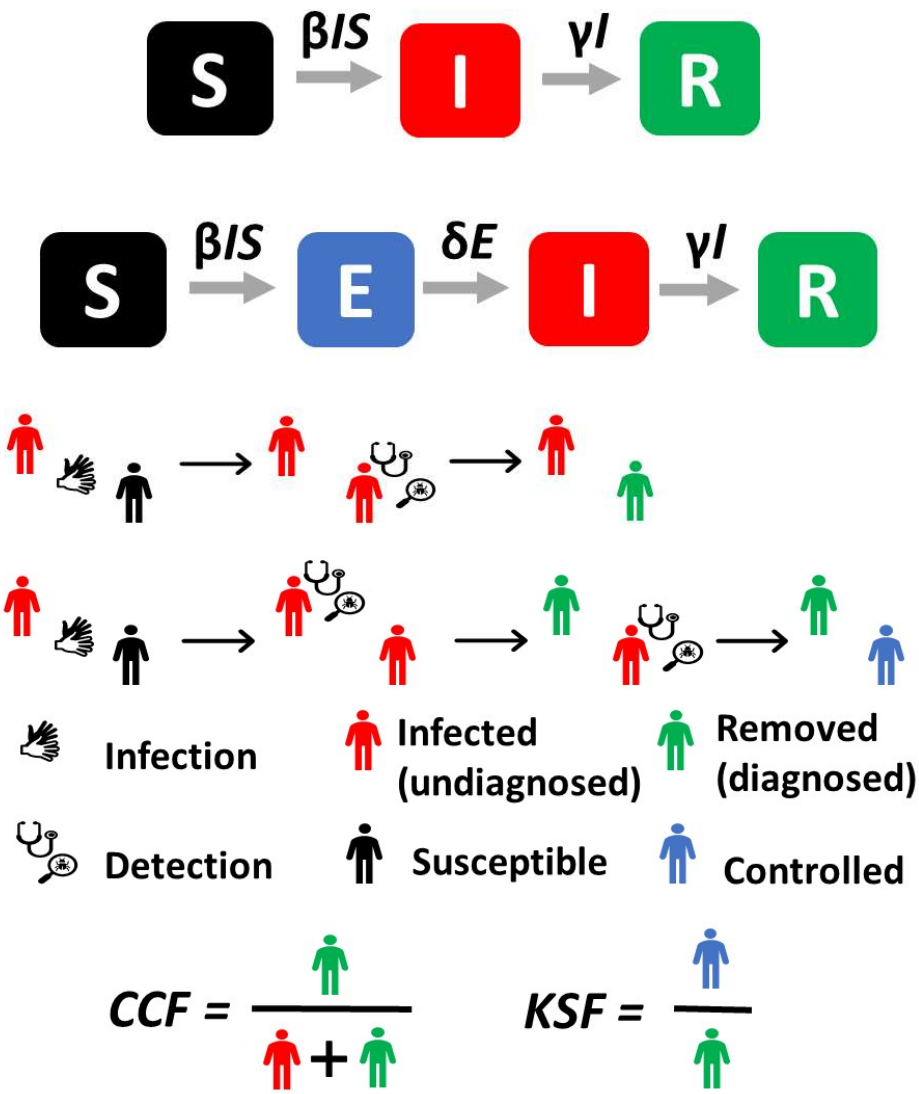

Figure 1: Models Description
Upper plot - Dynamics of the SIR model. A Susceptible individual can get infected with a rate proportional to βIS. An infected can get Removed from the system with a rate proportional to γI. Middle plot - Dynamics of the SEIR model. An Exposed category is added. Exposed are not infecting but can become infecting with a probability of δ per exposed. Lower plot - Dynamics of the discrete-time simulations: First line - Each Infected (red) can infect each susceptible (dark). If an infected is detected, it becomes quarantined and thus removed (green). Second-line If the contaminator of a diagnosed individual was already detected (i.e it is green by the time the new infected is diagnosed), the newly diagnosed is considered "controlled" (blue) implying that its source of contamination is known. We define two ratios. *CCF* is the fraction of diagnosed individuals over the total number of infected (diagnosed and undiagnosed). *KSF* is the fraction of diagnosed individual with a known source of contamination.

Methods

We performed discrete stochastic simulations of both SIR and SEIR models for different infectivity distributions, where each event is explicitly modeled. The models studied either had an equal probability of getting infected for each susceptible, or a

variable distribution with a scale-free distribution. We present here results with a slope of -2, but other slopes had similar results.

Following is a technical description of the simulation framework. For the sake of efficiency, each event (e.g. infection, detection…) is represented as a tree to allow a rapid selection of the individual involved in the next event. Each leaf corresponds to an individual. The value of each internal node in the tree is the sum of the values in its direct descendants in the tree. The tree root is the total probability of the event. This configuration enables us to access each individual in logarithmic time. We also keep track of the identity of the contaminator in a repertoire, in case of a contamination event.

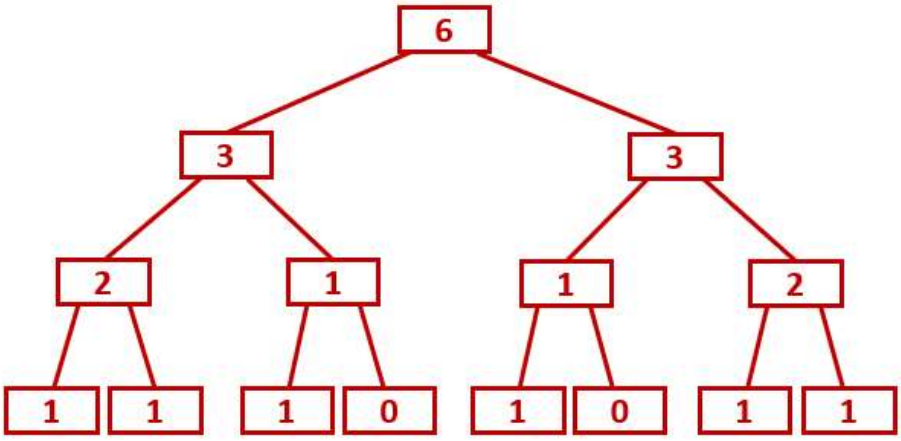

Figure 2: Simulations methodology
Trees recording individuals in a category of the population. There are trees for each event. Each node is the sum of its two direct descendent. The leaves are events involving a given individual (from the appropriate category). The root is the total probability of an event in the entire population. The leaves can have different values if the probability of an event differs among individuals.

We compute the normalized probabilities of each event (based on the top node of the tree of this event) in the appropriate model. At each step, we choose an event based on these probabilities. For a contamination event, a susceptible is chosen based on its (pre-defined) infectivity. The probability of such an event is the product of the total number of infected, the total infection probability of susceptible individuals, and the infection rate. Following, an infection event, a susceptible becomes infected, the chosen susceptible is determined by traversing the susceptible tree. The tree is then updated along the entire path. We also choose randomly an infected as the contaminator and record its leaf number in the repertoire.

For a detection event, an individual is randomly chosen in the infected tree with a probability proportional to the product of the total number of infected and the detection rate. We then check if his/her contaminator has already been detected by observing if the leaf of the contaminator was already detected. In such a case, the number of Controlled is increased by 1. Once the total number of infected reaches one percent of the total population, we stop the simulation. The ratios in figures 3A and B are taken along the simulations. The results in figure 3C are at the last time point of the simulation. Simulations where the total number of infected collapsed before reaching one percent of the total population were not incorporated in the results.

Results

In many cases, the *KSF* can be estimated from the epidemiological investigations even on a limited sample of the confirmed infected individuals. In contrast, *CCF* can be directly measured only through wide scales surveys. Moreover, in most countries, the total fraction of infected is low, requiring very large surveys to obtain accurate estimates.

To test whether *KSF* can be used to estimate *CCF*, we computed *KSF* and *CCF* in different models. We then tested whether the relation between *KSF* and *CCF* is sensitive to the details of the model. While there is a large number of existing models, most current works on COVID-19 are based on different versions of either the SIR or SEIR models.[4,19,22] We thus simulated these usual three (four) categories: Susceptible (*S*), (Exposed (*E*)), Infected/undiagnosed infected (*I*), and Removed/Diagnosed (*R*). The last category includes all individuals that can no longer contaminate others, namely all the diagnosed individuals. Therefore, the Infected category corresponds to undiagnosed individuals. This is a reasonable assumption since in most countries diagnosed individuals are put into quarantine. We also assume that recovered individuals can no longer get infected and, therefore, do not return to the Susceptible category (at least not with a high enough probability). β represents the infection rate at which Susceptible get Infected (Exposed), (δ is the rate at which Exposed become Infectious), and γ is the detection rate at which Infected get Removed. We further define as *C* the number of Controlled individuals. It represents the fraction of the Removed for whom the identity of their contaminator is known. In practice, each time a Susceptible gets infected, an Infected is randomly chosen to be the contaminator and its identity is recorded. When an individual gets diagnosed, we check the identity of its contaminator and if this contaminator has already been diagnosed, we consider that the newly diagnosed individual moves to the Controlled category (see figure 1 for a description of the dynamics and Methods). We ignored false positives (diagnosed that are not infected) in the current analysis, as their number is consistently small.[17,26] We further discuss false negatives.

To simulate epidemics, we keep track of *I*, *R* and *C* and compute the two ratios: $\frac{1}{CCF} = 1 + \frac{I}{R}$ and $\frac{1}{KSF} = \frac{R}{C}$. Those two ratios rapidly achieve equilibrium (plots A and B in figure 3). Moreover, in different realizations of the same model, most of the trajectory density is centered on a limited range of *KSF* and *CCF* values. Different initial conditions and stochastic realizations lead to similar solutions (figure 3B). As such, for a given model and known parameters, one can use *KSF* to estimate *CCF*.

However, in most cases, the spread dynamics parameters or even the appropriate model are unknown. To show that the relation between *KSF* and *CCF* is not model or parameter specific, we tested this relation in multiple versions of SIR and SEIR models and different parameter configurations. We implemented SIR and SEIR models with homogenous and heterogeneous infection rates to reflect the fact that not all individuals have the same infection probability (as a function of age/gender/genetics or other factors). In each configuration, we ran simulations with different values of the parameters (figure 3C). One can see that, while both fractions vary in different models and parameters, however, an approximately linear relation is consistent among all models.

Since *R* and *C* can be obtained from measures of diagnosed infected and epidemiological investigations, *KSF* can be estimated in most cases. Then *CCF* and thus *I* can be determined from the relation in figure 3C. To check the applicability of our methodology, we analyzed the number of confirmed cases in Israel every day[7]. In parallel, we analyzed from the Israeli Minister of Health the fraction of confirmed cases in Israel with a known source (*KSF*) (figure 3D). We then estimated the total number of infected in Israel (figure 3E). Note that in Israel, the number of undiagnosed infected is relatively small since this country has applied stringent controls and testing early in the crisis. This agrees with the relatively low number of death events in Israel.

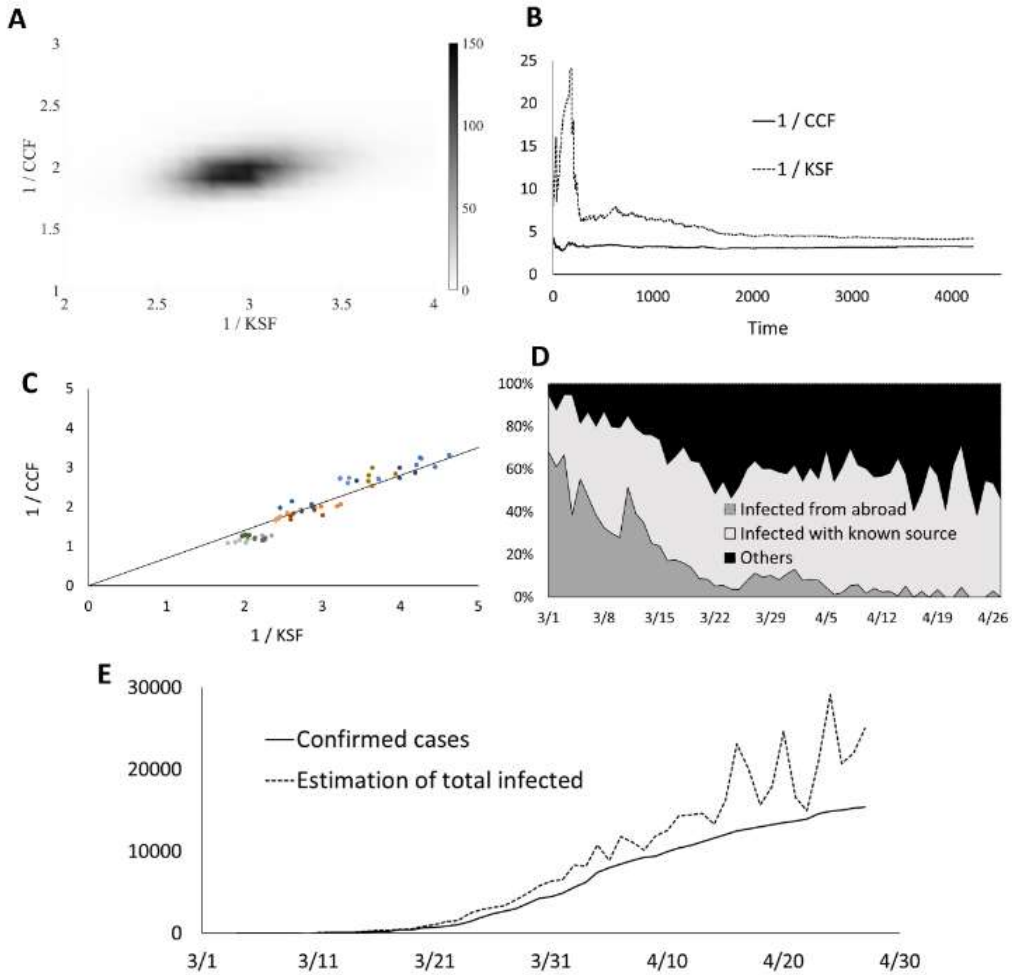

Figure 3: Results from simulations
Plot A - Density plot of the two ratios over 1000 simulations. Plot B - Time evolution of the two ratios $\frac{1}{CCF} = 1 + \frac{I}{R}$ and $\frac{1}{KSF} = \frac{R}{C}$. We observe that not only the ratios achieve equilibrium, but they are never very far from it. Plot C - Independently of the model used (SIR or SEIR, with homogeneous or heterogeneous infection rate), we observe a linear relation between *1/CCF* and *1/KSF*. Plot D – Distribution of source of infection per day in Israel. Data. obtained from the Israeli ministry of health with the fraction of confirmed cases with a known source. We ignored in this analysis infected coming from abroad (deep gray). Plot E - The number of confirmed cases (Removed) in Israel was obtained from world data (full line).[7] We used our method to estimate the total number of infected in Israel (Dashed line).

Discussion

We have presented a method to estimate the fraction of undiagnosed infected from the fraction of infected with a known contaminator (out of all infected). While the first value is hard to measure in realistic situations, the second is often known.

The *KSF* estimate suffers from multiple caveats with opposite effects. First, removed individuals were considered controlled only if their contaminator was already diagnosed when in fact it could be diagnosed even after. Therefore, even already removed individuals could be counted eventually as controlled. A second and more complex problem is that reported infected may be biased toward people who have been in contact with other reported infected. As such, the number of controlled individuals would be overestimated. A direct solution to these limitations would be to perform detailed epidemiological investigations on patients with clinical complications. Such patients typically do not suffer from sampling bias and detailed enough investigations will limit the number of missed controls. Such investigations can be performed on a limited sample.[27]

Another limitation of our estimate is that epidemiological investigations are not perfect, as such, some controlled individuals might be missed. Similarly, some diagnosed may be assumed to be infected from a known source, when in fact they were infected by other sources. These limitations can be solved when detailed genetic information is available on the virus. Note again that only a small fraction of the diagnosed individuals has to be investigated in detail to obtain *KSF*.

As of May 31st, the number of daily infections is decreasing in most countries, but this decrease is slower than the predictions of all SIR models class. This feature comes from the fact that the infection rate, although lower than at the beginning of the crisis, is still high and erratic. If new models were to be developed for such periods, the same analysis can be done to test whether their results stay on the same approximate linear relation between *1/CCF* and *1/KSF* even if SIR style models are abandoned.

Other versions of the SIR models include a transition to a death state or an Asymptomatic category. Since our Removed category includes all individuals that can no longer contaminate, it already accounts for the dead and the effect of quarantine. Our Infected category includes all undiagnosed individuals that can contaminate others therefore, it accounts for all carriers including the asymptomatic individuals. Migration has a minor effect on contamination,[23] so we did not include this feature. For the sake of simplicity, we presented here a non-spatial model where all infected individuals can contaminate others disregarding proximity but, since the similarity between *CCF* and *KSF* is an inherent property of epidemiological models, we do not expect network and spatial features to change our conclusions.

To summarize, as is the case for every model, multiple caveats can affect the validity of the model, most of those can be avoided in detailed and unbiased investigations on small numbers of diagnosed (even a few tens). Thus, while we do not propose to use the observed relation as is on biased published epidemiological data, the here reported relation between *KSF* and *CCF* can be a critical tool to estimate the spread of diseases.

References

1 Prem K, Liu Y, Russell TW, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health* 2020; **5**: e261—70.

2 Tian H, Liu Y, Li Y, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* 2020; **368**: 638—42.

3 Yuen KS, Ye ZW, Fung SY, Chan CP, Jin DY. SARS-CoV-2 and COVID-19: The most important research questions. *Cell & bioscience* 2020; **10**: 1—5.

4 Grant A. Dynamics of COVID-19 epidemics: SEIR models underestimate peak infection rates and overestimate epidemic duration. *medRxiv* 2020.

5 Richterich P. Severe underestimation of COVID-19 case numbers: effect of epidemic growth rate and test restrictions. *medRxiv* 2020.

6 Theagarajan LN. Group testing for COVID-19: how to stop worrying and test more. *arXiv* 2020.

7 https://ourworldindata.org/coronavirus#testing-for-covid-19

8 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** :689-97.

9 Goswami G, Prasad J, Dhuria M. Extracting the effective contact rate of COVID-19 pandemic. *arXiv* 2020.

10 Verity R, Okell LC, Dorigatti I, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet infectious diseases* 2020; **20**: 669—677.

11 Kucharski AJ, Russell TW, Diamond C, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet infectious diseases* 2020; **20**: 553—58.

12 Walker PG, Whittaker C, Watson O, et al. The global impact of COVID-19 and strategies for mitigation and suppression. *Imperial College London* 2020.

13 Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epidemic processes in complex networks. Rev. Mod. Phys. 2015; **87**: 925.

14 Yanev NM, Stoimenova VK, Atanasov DV. Stochastic modeling and estimation of COVID-19 population dynamics. *arXiv* 2020.

15 Flaxman S, Mishra S, Gandy A, et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. *arXiv* 2020.

16 Zhigljavsky A, Whitaker R, Fesenko I, et al. Generic probabilistic modelling and non-homogeneity issues for the UK epidemic of COVID-19. *arXiv* 2020.

17 Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, Wang M. Presumed asymptomatic carrier transmission of COVID-19. *Jama* 2020; **323**: 1406—7.

18 Wearing HJ, Rohani P, Keeling MJ. Appropriate models for the management of infectious diseases. *PLoS medicine* 2005; **2**: 174.

19 Giudici M, Comunian A, Gaburro R. Inversion of a SIR-based model: a critical analysis about the application to COVID-19 epidemic. *arXiv* 2020.

20 Xu L, Zhang H, Deng Y, et al. Cost-effectiveness Analysis of Antiepidemic Policies and Global Situation Assessment of COVID-19. *arXiv* 2020.

21 Loli Piccolomini E, Zama F. Preliminary analysis of COVID-19 spread in Italy with an adaptive SEIRD model. *arXiv* 2020.

22 Pandey G, Chaudhary P, Gupta R, Pal S. SEIR and Regression Model based COVID-19 outbreak predictions in India. *arXiv* 2020.

23 Chinazzi M, Davis JT, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* 2020; **368**: 395—400.

24 Singh R, Adhikari R. Age-structured impact of social distancing on the COVID-19 epidemic in India. *arXiv* 2020.

25 Berger DW, Herkenhoff KF, Mongey S. An seir infectious disease model with testing and conditional quarantine. *National Bureau of Economic Research* 2020.

26 Yan G, Lee CK, Lam LT, et al. Covert COVID-19 and false-positive dengue serology in Singapore. *Lancet Infectious Diseases* 2020; **20**: 536.

27 Mueller M, Derlet PM, Mudry C, Aeppli G. Using random testing to manage a safe exit from the COVID-19 lockdown. *arXiv* 2020.

28 Weitz JS, Beckett SJ, Coenen AR, et al. Modeling shield immunity to reduce COVID-19 epidemic spread. *Nature Medicine* 2020; **7**: 1—6.

29 Winter AK, Hegde ST. The important role of serology for COVID-19 control. *The Lancet Infectious Diseases*.