Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company’s public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A phenomenological estimate of the true scale of CoViD-19 from primary data

Luigi Palatella\textsuperscript{a}, Fabio Vanni\textsuperscript{b,c,d,e}, David Lambert\textsuperscript{e,f}

\textsuperscript{a} Liceo Scientifico Statale “C. De Giorgi”, Lecce, Italy
\textsuperscript{b} Sciences Po, OFCE, France
\textsuperscript{c} Institute of Economics, Sant’Anna School of Advanced Studies, Pisa, Italy
\textsuperscript{d} Center for Nonlinear Science, University of North Texas, USA
\textsuperscript{e} Department of Mathematics, University of North Texas, USA
\textsuperscript{f} Department of Physics, University of North Texas, USA

\textbf{A R T I C L E   I N F O}

Article history:
Received 24 January 2021
Revised 27 February 2021
Accepted 3 March 2021
Available online 14 March 2021

Keywords:
Renewal equation
Scale of epidemics
SARS-CoV-2 prevalence

\textbf{A B S T R A C T}

Estimation of the prevalence of undocumented SARS-CoV-2 infections is critical for understanding the overall impact of CoViD-19, and for implementing effective public policy intervention strategies. We discuss a simple yet effective approach to estimate the true number of people infected by SARS-CoV-2, using raw epidemiological data reported by official health institutions in the largest EU countries and the USA.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

As the coronavirus disease 2019 (CoViD-19) epidemic reached every corner of the world, each country has adopted different interventions to manage the early and long-term phases of the spread of the epidemic. This epidemic has forced many countries to react by imposing policies aimed at reducing population mobility together with internal and international border limitations. These policies have been informed by various measures of epidemic risk that all trace back to the number of new cases that the country’s healthcare system has found each day. Unrecorded and unnoticed cases make it difficult to estimate the true number of infected persons present at a given time (prevalence).

Cases of infection are usually detected through testing, typically occurring when ill people (or their recent contacts) seek healthcare. Official data are mainly collected with medical swabs. This favors the examination of patients showing clear symptoms. However, key features of CoViD-19’s dynamics have to do with asymptomatic or pre-symptomatic transmission \cite{1,2}. Transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) can occur before symptom onset in the infector, which presents a stumbling block to efforts to stop the spread of the disease. Infected persons often do not develop noticeable symptoms until after the viral latent period, the time between being infected with the virus and first becoming contagious. (This differs from the incubation period: the time between being infected with the virus and first developing symptoms.) Recent studies \cite{3–5} have found that SARS-CoV-2 has the notable property that its latent period is shorter than the incubation period of the virus. Thus, individuals can become contagious even before they show symptoms. Early in the epidemic, and in various situations afterward, diagnosis has been based on having a certain set of symptoms. More commonly, an infection diagnosis comes from direct detection of SARS-CoV-2 RNA by nucleic acid amplification tests, typically reverse-transcription polymerase chain reaction (RT-PCR) from the upper respiratory tract. Furthermore, testing capacity depends on the demands a healthcare system is under. So that the confirmed case numbers reported during an outbreak represent only a fraction of the true levels of infection in a community.

Undocumented infections often are not detected due to mildness, limitedness, or absence of symptoms. As they are not quarantined, they can expose far more of the population to the virus, thereby sustaining the spread of the epidemic. Moreover, the fraction of undocumented (but still infectious) cases is a critical epidemiological feature that modulates the pandemic potential of a virus. The ability to estimate the scale of an epidemic is of paramount importance medically, socially, and economically, as it affects viral hot-spot detection, resource allocation, and intervention planning. Restrictive measures have indeed exempted firms...
producing essential goods and services. In addition, companies able to massively employ remote work have succeeded in mitigating the negative effects. Better estimation of the true prevalence of COVID-19 cases is crucially important to set the strength and scale of non-medical interventions and improve the evaluation of the economic and health impacts associated with lockdown and reopening policies.

Many modeling methods use the mortality rate to estimate the scale of COVID-19 [6–9]. These methods rely on fixed estimates of certain epidemiological properties (such as the onset-to-death interval distribution and the generation-time distribution). Further, they assume that the fatality rate is the same everywhere and that its true value is near the lowest observed value (\(\approx 0.5\text{–}1.0\%\)). Finally, these methods require a long time to produce results. We propose an independent way to estimate the true incidence of SARS-CoV-2. The intuition behind this approach lies in the well-known fact that depletion of the susceptible population reduces the reproduction number \(R_0\). We make a quantitative analysis of this saturation effect to estimate the true scale of the COVID-19 pandemic, i.e., the order of magnitude of the actual number of people that have been infected.

In the Methods section, we explain the mathematical background for the estimation of the fraction of un-diagnosed cases by using both the number of new positive tests and the instantaneous reproduction number \(R_t\). In the Data and Results section, we show the results for Italy, France, Spain, Germany, and USA. In the Conclusions section, we summarized our research and conclude the paper.

2. Methods

At the very beginning of an epidemic, the fraction of the population that is susceptible to the disease is 1. The basic reproduction number, \(R_0\), is the average number of people that are infected by a single infected person per day when everyone is susceptible. In general, \(R_0\) depends on the capacity of the virus to infect people, as well as on aspects that vary from country to country such as social habits, movement patterns, average health status, and sanitation. Moreover, the reproduction number may depend on time due to seasonal conditions, air humidity, UV exposure, and social changes (e.g., an increase in mask wearing or increased average interpersonal distance). When people become immune or die the susceptible fraction of the population decreases. All else being equal, this leads to a decrease in the instantaneous (or apparent, or actual) reproduction number, \(R_t\). This phenomenon is described by the following relation:

\[
R_t = R_0(t)(1 - C(t)).
\]

Here, \(C(t)\) is the true fraction of the population that has been infected at time \(t\). Let us suppose that the testing system of a region is not able to detect every positive case, we define

\[
\lambda(t) \equiv \frac{\text{detected cases}}{\text{actual cases}} = \frac{c(t)}{C(t)}.
\]

where \(c(t)\) is the total number of detected cases (the cumulative number of cases) as a fraction of the whole population.

The instantaneous reproduction number, \(R_t\), is an important tool for detecting changes in disease transmission over time [10,11]. Policy makers and public health officials use \(R_t\) to assess the effectiveness of interventions and to inform policy. Specifically, the instantaneous reproduction number measures disease transmission at a given point in time, \(t\). One can interpret it as the average number of people that each contagious individual at time \(t\) would infect per day, if conditions remained unchanged. We assume, following [12–14], that \(R_t\) depends on the number of new cases per day, \(j(t)\), only through \(1 - C(t)\) (as in Eq. (1)) and that \(j(t)\) obeys the so-called renewal equation:

\[
j(t) = \frac{d}{dt}C(t) = \int_0^\infty A(t, \tau)j(t - \tau)d\tau + i(t),
\]

where \(i(t)\) is the number of imported cases per day, and \(\tau\) represents the infection age, the amount of time since an individual became infected.

The function \(A(t, \tau)\) is a product of \(R_t\) and the generation-time distribution \(g(t, \tau)\) [15]. The distribution \(g(t, \tau)\) represents the probability density for a person infected at time \(t - \tau\) to be infectious at time \(t\) (as with any probability density \(\int_0^\infty g(t, \tau)d\tau = 1\)). For simplicity, we assume that this distribution is independent of \(t\), i.e., \(g(t, \tau) = g(\tau)\). So that

\[
j(t) = R_t\int_0^\infty g(\tau)j(t - \tau)d\tau.
\]

Typically, the generation distribution is unknown, though it can be approximated by assuming it is the same as the serial-interval distribution, which refers to the time between successive cases in a chain of transmission. We follow [16] in using a gamma function given by

\[
g(\tau) = \beta^\alpha\tau^{\alpha - 1}e^{(-\beta\tau)}
\]

with \(\alpha = 1.87\) and \(\beta = 0.28\). We estimate \(R_t\) by [10]

\[
R_t = \frac{j(t)}{\int_0^\infty g(\tau)j(t - \tau)d\tau} \approx \frac{j(t)}{\int g(\tau)j(t - \tau)d\tau}.
\]

where \(j(t) \equiv \frac{dC}{dt} = \frac{d\lambda(t)C(t) + \lambda(t)j(t)}{d\tau}\). We assume that \(\lambda(t)\) changes slowly enough that \(j(t) \approx \lambda(t)j(t)\) and \(\lambda(t)\frac{dt}{d\tau} = 1\) (for \(0 \leq \tau \leq \frac{1}{\beta}\)), so that the approximation in Eq. (6) holds.

Starting from the saturation effect of Eq. (1), we wish to find two unknown variables namely, \(R_0\) and \(\lambda\). With ideal data, one could immediately apply linear fitting techniques. However, the data reported by most countries are affected by a strong weekly fluctuation. To reduce this effect, we perform a moving average of the \(R_t\) obtained over 7 or 14 days (the results are almost identical).

As Fig. 2 shows, \(R_0(t)\) abruptly increased at the end of September, then decreased in the first days of November. Since \(R_0(t)\) varied significantly, a simple linear fit with \(R_0\) as a parameter would yield poor results. Instead we make an independent estimate of \(R_0(t)\) by doing a weighted average of the \(R_t\’s\) obtained in all the regions in a country at a given calendar time \(t\). This estimate reads as:

\[
(R_0(t)) = \frac{1}{N_{\text{regions}}} \sum_{k \in \text{regions}} R_t^{(k)} 1 - c^{(k)}(t).
\]

where \(c^{(k)}(t)\) is the fraction of the population recorded as having been infected in region \(k\) up to time \(t\), while \(R_t^{(k)}\) is the value obtained using Eq. (6) at time \(t\) in region \(k\).1 We apply a linear fit to \(R_t/(R_0(t))\) as a function of \(c(t)\). Inserting Eq. (7) into Eq. (1), we obtain

\[
\frac{R_t}{(R_0(t))} = (1 - C(t)) \approx \left(1 - \frac{c(t)}{\lambda}\right).
\]

So we interpret the intercept of the fit as an extrapolation of the initial susceptible population fraction and the slope as \(\frac{1}{\lambda}\).

---

1 One should use instead \((R_0(t)) = \sum_{k \in \text{regions}} R_t^{(k)} 1 - c^{(k)}(t)\); but the value of \(\lambda\) at this stage is unknown. We use a recursive approach inserting the value obtained in the linear fit procedure explained below and then repeating the whole procedure. In any case we see a negligible difference in the value of \(\lambda\) obtained by using the simplest factor \(1 - c^{(k)}(t)\) as in Eq. (7), so we present the simplest approach.
In this section, we apply our method to data and obtain satisfactory results. Our data sources are listed in Table 1 below.

The analysis was performed at the regional level in all the countries, specifically: Regione for Italy, Region for France, Länder for Germany, Comunidad autónoma for Spain, Federal States for the USA, and Regions of England plus Wales, Scotland, and Northern Ireland for the UK. For each region, the total population used to calculate \( c(t) \) was retrieved from the most recent census available online (except for regions of France which report population size directly in the dataset). All the data files offer extended epidemiological reports; however, we only used the number new Covid-19 cases per day.

In some cases, official institutions supply data by date of onset of symptoms, as in the case of Italy, see Fig. 1. This type of data is especially nice for finding the instantaneous reproduction number \( R_t \). Indeed, Italian Istituto Superiore di Sanità currently uses these values because they are virtually unaffected by fluctuations in testing capacity. Unfortunately, these data are only available for the whole country rather than particular regions. So, we only used these data to compare our measure of \( R_t \) with the one based on the raw data of new infections. We find that our measure is accurate up to reasonable statistical fluctuations.

Now we turn to Figs. 2 and 3. For each country we plot the value of \( R_t^{(k)} \) for its regions versus time and \( R_t / (R_0 t) \) versus \( c(t) \) along with the best fit for \( \lambda \). The fit was performed neglecting the first 100 days of the epidemic when testing was quite irregular and the number of PCR tests extremely low. The error on \( \lambda \) was evaluated by changing this threshold from 50 to 150 days.

A straightforward application of the estimate of the true number of infections is to adjust the case fatality rate (CFR) to better approximate the infectious fatality rate (IFR). The CFR is defined as the number of reported deaths per number of reported cases. It approximates IFR, which is an estimate of the death rate among all those infected with SARS-CoV-2, the virus that causes Covid-19. In Table 2 we estimate the unrecorded fraction of the true number of infected individuals as well as the corrected value of the case fatality rate (\( \lambda \)-CFR). This correction does not capture every reason the CFR can differ from the IFR. Some additional differences are due to delayed effects, population age, typical medical conditions of the population, and healthcare system efficiency.

As the vaccination campaign continues, we must consider the effects of immunization through vaccination on our analysis. The functional form will change to \( R_t / (R_0) \sim (1 - c(t)/\lambda)(1 - \nu(t)) \sim (1 - c(t)/\lambda - \nu(t)) \), where new term takes in account the vaccinated fraction of the population \( \nu(t) \). If data on the vaccination campaign are available at the regional level, and \( \nu(t) \) reaches a substantial value, one should perform a two-dimensional fit of \( R_t / (R_0) \) as a function of \( c(t) \) and \( \nu(t) \). The slope of the iso-\( R_t / (R_0) \) lines, i.e., the lines in the \( c(t) - \nu(t) \) plane of constant \( R_t / (R_0) \), will give another estimate of \( \lambda \). This will allow us to compare and contrast the effects of vaccination and new infectious cases on \( R_t \). If \( \lambda \approx 1 \), their effects should be the same. If \( \lambda < 1 \), we expect that the effect of a given number of infectious individuals on \( R_t \) to be larger than that of the same number of vaccinated.

### 3.1. Robustness check using seroprevalence data

Our findings can be checked by comparing our results with alternative methods used to evaluate the true number of people
that have had SARS-CoV-2. One of the most reliable methods is the analysis of seroprevalence of IgG antibodies in blood. We compare our λ values with those obtained by two papers [26,27] and a dashboard on an official website [28].

In particular, Havers et al. [26] reports the fraction of people showing IgG antibodies to SARS-CoV-2 over the number of officially reported cases as of the end of May 2020. This ratio should be compared with λ−1. In Table 3 of [26] data for 10 sites are reported. They range from a minimum of 6.0 for Connecticut to a maximum of 23.8 for Missouri. The worst hit state, New York, has a value of 11.9. In Fig. 4 (left panel), we report the value of λ using data covering the same period, namely ending May 23rd, 2020. We report our best fit yielding λ = 0.06, together with the line corresponding to the λ obtained from serological data described in Havers et al. [26]. The same approach was followed in Fig. 4 (right panel), where we report the typical value retrieved from [28].
(between 6 and 7 times the official number of cases) with our fit at the date of the last update of the website (July 2020).

Passing from late spring to the end of summer, \( \lambda \) increased considerably. This effect is apparent in the dashboard data [28] and the same pattern holds in all countries analyzed. We presume this increase is due to increased testing ability and contact tracing efforts made in all countries. Nevertheless, as a final remark, we stress that if our results are confirmed, in all countries the tracing capacity is not enough, by itself, to mitigate the spread of SARS-CoV-2. Due to the large contribution of asymptomatic or mild-symptomatic cases, we think that reaching a value of \( \lambda \simeq 1 \) would be necessary. This would require a massive use of tests as seen, for example, in China during its prevention of a second wave.

Finally, in comparing our approach to the seroprevalence data, we found that our method gives results that are of the same order of magnitude. Our method tends to overestimate the serological results. This could be mark against our approach, but one should note that seroprevalence-based infection estimates are conserva-
tive [28]. Some studies have found that infected persons who are asymptomatic or have mild symptoms do not have detectable antibodies. Other studies found that antibody levels decrease and become undetectable in some patients over time (see, for example, Le Bert et al. [29], Ng et al. [30]). Another explanation of this over-estimation is that the number of susceptible individuals is not the entire population due to cross-immunity with other common coronaviruses [31–33].

4. Conclusions

Knowledge of true prevalence of CoViD-19 is critical for informing policy decisions about how to distribute resources and manage the impacts of CoViD-19 on public health, society and the economy [34,35]. The true scale of the epidemic can affect economic development since it reduces long-run economic growth by limiting the size of social networks. On the contrary a low prevalence estimate could lead people to take more epidemic-relevant risks, making disease eradication impossible using social distancing policies only.

We have proposed a method to estimate the true number of infected people, unveiling the true scale of CoViD-19 by using PCR test data alone. We, then, have used this method to estimate a more reliable case fatality rate. Our approach is a phenomenological estimate of the true scale of the epidemic, since it is based on an empirical relationship between phenomena, in a way which is consistent with fundamental theory, but is not directly derived from that theory. Consequently, our method can be affected by errors which we are not able to distinguish from the lack of theoretical understanding. Nevertheless, there are two remarkable results. The few attempts to assess the true impact of SARS-CoV-2 via serological tests yield λ-values comparable in order of magnitude with our calculations [36]. Moreover, countries that performed better tracing, like the USA, experienced a fatality rate lower than Italy, France, and Spain (in the first wave). Improved contact tracing performance (higher λ) leads to a lower fatality rate because the rate is calculated using the correct denominator. Good contact tracing can also help keep the epidemic under control, preventing the virus from reaching fragile and elderly people. Thus leading to a bona fide decrease in the fatality rate.

We have proposed that the study of the graph R₁ vs. c(t) and R₁/(R₀(t)) vs. c(t) can provide useful insight into epidemic dynamics. These methods should be useful in other epidemics with incomplete data due to the presence of unrecorded contagious individuals, for example, the seasonal flu. Minor modifications allow this approach to keep its effectiveness even in the presence of active vaccination campaigns. By contributing to a better-informed response to epidemics, we believe this work serves to benefit society as a whole.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

CRediT authorship contribution statement

Luigi Palatella: Conceptualization, Formal analysis, Data curation, Writing - original draft, Writing - review & editing.
Fabio Vanni: Conceptualization, Formal analysis, Data curation, Writing - original draft, Writing - review & editing.
David Lambert: Conceptualization, Formal analysis, Data curation, Writing - original draft, Writing - review & editing.

Acknowledgments

Fabio Vanni acknowledges support from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 822781 GROWNPRO - Growth Welfare Innovation Productivity.

References

[1] Meyerowitz EA, Richterman A, Bogoch II, Low N, Cevik M. Towards an accurate and systematic characterization of persistently asymptomatic infection with SARS-CoV-2. Lancet Infect Dis 2020. doi:10.1016/S1473-3099(20)30837-9.
[2] Moghadas SM, Fitzpatrick MC, Sah P, Pandey A, Shoukat A, Singer BH, et al. The implications of silent transmission for the control of CoViD-19 outbreaks. Proc Natl Acad Sci 2020;117(30):17515–15.
[3] Peirlinck M, Linka K, Costabal FS, Kuhl E. Outbreak dynamics of CoViD-19 in China and the United States. Biomech Model Mechanobiol 2020;19(6):2179–93.
[4] Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of coronavirus disease 2019 (CoViD-19) from publicly reported confirmed cases: estimation and application. Ann Intern Med 2020;172(9):577–82.
[5] Byrne AW, McEvoy D, Collins AB, Hunt K, Casey M, Barber A, et al. Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic CoViD-19 cases. BMJ Open 2020;10(8). https://bmjopen.bmj.com/content/10/8/039856.full.pdf.
