Uncovering the universality of COVID-19 fatality rates

João Rodrigo Souza Leão[a], *Gustavo Zampier dos Santos Lima[a]

[a] *Universidade Federal do Rio Grande do Norte, School of Science and Technology (ECT), Natal-RN, 59078-970, Brazil*

Abstract

An important and elusive characteristic of the COVID-19 pandemic is the fatality rate which allows us to understand the severity of this disease, health care needs, and the impact on large populations. To address this and other questions we present a probabilistic model to study the evolution of the COVID-19 pandemic and to correct the case fatality rates. Our model employs probabilities to estimate the time evolution of infections, recoveries, and deaths. This model discriminates asymptomatic, mild/moderate, and severe cases and allows for the estimation of undiagnosed individuals. Furthermore, we compare the model curves to official data for medium-sized cities, world metropolises, and medium-sized countries, spanning a range of populations from a few million to several million individuals. Using the undiagnosed estimates we correct the case fatality rates and find that it ranges from $\mathbf{0.33\% \pm 0.02\%}$ to $\mathbf{1.14\% \pm 0.07\%}$. Since we applied the method to cities and countries with different characteristics, the corrected case fatality rates indicate a universality that is independent of location and other social/demographic conditions. Our results agree with sample tests and seroprevalence studies, considerably changing our understanding of the COVID-19 fatality rates. Other applications include the estimate for severe cases, ICU needs, and undiagnosed cases.

Keywords:
COVID-19 pandemic, Corrected Case Fatality Rate, Probabilistic Model, ICU needs, undiagnosed cases

Email addresses: `rodrigoleao@ect.ufrn.br` (João Rodrigo Souza Leão), `guzampier76@gmail.com` (*Gustavo Zampier dos Santos Lima)

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS COV 2) is the virus causing the Corona Virus Disease (COVID-19) [1]. Sources suggest the virus crossed from their animal hosts to humans through an yet unknown mechanism [2, 3, 4]. It started infecting humans in late 2019 in Wuhan, China and by March 2020 the disease was spreading fast in several countries around the globe and was declared a global pandemic by the World Health Organization (WHO) [5, 6, 7]. The SARS-COV-2 virus is known to cause a wide variety of symptoms, most notably in the respiratory system [8, 9] and it spreads through human contact, contaminated surfaces, and through the air, [10]. This latter mechanism is particularly important if people are in confined air-tight spaces such as airplanes, restaurants, and similar environments. As in other pandemic episodes, isolation plays a major role in containing the virus [11, 12]. Very early in the pandemic, the governments imposed social isolation, confinement, and even complete lock downs lasting several weeks in an effort to stop the virus dissemination. This generated immense political, social, and economic burdens with consequences virtually impossible to measure at the present time [13, 14, 15]. Health professionals and governments worldwide are making huge efforts to understand the virus, treat the patients and at the same time keep track of infected individuals [16].

One of the fronts in the battle to understand epidemics and to plan effective control strategies is guarded by mathematical models [17, 18, 19, 20, 21, 22]. These so-called dynamic or epidemiological models try to answer some general questions: What is the duration of the epidemic?; How many people will get infected and die?; Are isolation and social distancing determinants to the course of the pandemic? The answers to these questions might help governments and health care officials to properly react. A widely used approach to model this problem is the Susceptible-Infectious-Recovered (SIR) model, first introduced in 1927 [23]. Other efforts use variations/generalizations of this approach to find exact and more general solutions [24, 25, 26]. Despite being limited and overly deterministic, SIR-like models continue to be useful and widely used to study pandemics and outbreaks [27].

In this work, we introduce a probabilistic model to study the COVID-19 pandemic. In particular, we assign relative probabilities and use random numbers to weight the occurrence of infections, deaths, and recoveries according to the assigned probabilities. This approach uses a straightforward and non-deterministic way to compute the time evolution of the disease and was specifically designed to model the characteristics of the COVID-19 pandemic. This model is especially useful to calculate the number of infected individuals, deaths, ICU needs, and especially the so-called asymptomatic cases that often go undiagnosed. One of the most important and elusive features of the COVID-19 pandemic is the case fatality rate (CFR) [28, 29]. Since asymptomatic cases are difficult to detect and might come in huge numbers, we anticipate that the number of cases is much larger than officially reported [30, 31, 32]. This implies that the case fatality rates need to be corrected for a proper estimation of the real fatality rate. Our model addresses this problem by dividing those infected into different levels of severity and specifically estimating asymptomatic cases that often go unreported.

This paper is organized as follows. In section 2 we fully describe our probabilistic model. Section 3 is devoted to discussing the parameter space of the model and fitting the model to real curves for selected cities and countries around the world. We also calculate the case fatality rates of the COVID-19 infection and correct them using the model results. Section 4 is dedicated to a thorough discussion of our results and an evaluation of the impact and reach of our probabilistic approach to modeling this problem. We also point out and discuss the limitations of the model. In section 5 we draw our conclusions, especially those regarding the corrections to the case fatality rates.

2. The Probabilistic Model

We developed a probabilistic code to model and interpret the time evolution of the COVID-19 pandemic. The model assumes that individuals may occupy different compartments according to their condition or stage of the disease: free, asymptomatic, mild/moderate, severe, recovered or deceased (Fig.1). The model works by assigning probabilities to these compartments and also to the related events (infections, recoveries, deaths). To weight these probabilities we use a very well-known and documented random number generating function [33]. For each day (step or iteration) the code evaluates if new infections, recoveries, or deaths occur in a completely random, non-deterministic fashion.

We aim to model a population of N individuals with a fraction I being isolated. First, we set the infection probability of free individuals (P_I) and assume that the probability of new infections varies according to the number of infected individuals (N_I) [34]. Since N_I varies with time it impĺies that P_I is also a function of time ($P_I(t)$). This happens because as the outbreak progresses more individuals became infected and the rate of new infections increases. This cannot go on indefinitely since both the total population N and the total number of exposed individuals $(1 - I) \times N$ are finite. It implies that as more individuals get infected, new infections are less likely, assuming re-infections do not occur. We thus define $P_I(t)$ as given by equation 1 thus allowing for a realistic modeling of the official cases:

$$P_I(t+1) = \begin{cases} P_I(t) + S & \text{if } N_I(t) \leq C_F \times N \\ P_I(t) & \text{if } C_F \times N < N_I(t) \leq (C_F + \delta) \times N \\ P_I(t) - S & \text{if } N_I(t) > (C_F + \delta) \times N \end{cases} \tag{1}$$

This function is updated at each iteration t (days) and increases by a constant value S (increase/decrease step). A critical fraction C_F of the total population ($C_F \times N$) is considered as a point of interest above which the function stops growing, in accordance with the idea of herd immunity since more recovered/immunized individuals decrease the rate of new infections. Thus the infection probability remains constant during a critical interval ranging from $C_F \times N < N_I(t) \leq (C_F + \delta) \times N$, where δ represents a small variation of the critical fraction C_F. In our simulations, we use $C_F = 0.40$ and $\delta = 0.10$, although these values can be reset by the user. When the critical interval is exceeded ($N_I(t) > (C_F + \delta) \times N$) the model assumes that new infections are rarer and thus the probability of infection starts to decrease with a negative S at each iteration. Eventually $P_I(t)$ is so low that the rate of new infections dims and the outbreak dies out.

Next we assign relative probabilities for each compartment in fig. 1. Free individuals, once infected, have a probability of assuming one of the infected compartments: asymptomatic (P_A), mild/moderate (P_M), or severe (P_S). In this model, only severe cases might evolve to death and we assign a probability P_D for this occurrence. Hence, the recovery probability of severe cases is $P_R = 1 - P_D$. The model has three levels of randomness and at each iteration, all individuals are checked. If an individual is free then the first aleatory number is drawn and compared against $P_I(t)$. If the individual becomes infected then a second aleatory number is drawn and compared to P_A, P_M, and P_S to check the severity of the case. Also at each iteration the model calculates/updates the number of asymptomatic ($A(t)$), mild/moderate $M(t)$, severe cases ($S(t)$), recoveries and deaths ($D(t)$). Asymptomatic and mild/moderate individuals always recover in t_A and t_M days after infection. Severe cases either recover or evolve to death in t_S days after infection when a third and final aleatory number is drawn and compared to P_D to decide if a death or a recovery will occur. It is very important to set these different recovery times to realistically model the official deaths and case curves. In our simulations we used fixed median recovery times (days) for the asymptomatic ($t_A = 10$), mild/moderate ($t_M = 15$) and severe cases ($t_S = 20$), according to [35, 36]. However, these values can be adjusted by the user if needed. As mentioned, a fraction I of the total population N is isolated/quarantined and the model will thus run on $(1 - I) \times N$ individuals. In summary, all the probabilities (P_A, P_M, P_S, P_D) and parameters (I, S, C_F, δ) are set by the user and reasonable values will be discussed in section 3.

It is known that the official number of total cases is often underestimated by a myriad of factors. Up to 75% of cases are asymptomatic [37, 38], meaning that these cases will rarely enter the official records. To properly model this feature we need to estimate a range for the number of cases to account for those that are not detected. Our model thus calculates both an upper limit ($N_I^{upp}(t) = A(t) + M(t) + S(t) + R(t) + D(t)$) and a lower limit ($N_I^{low}(t) = N_I^{upp}(t) - A(t)$) for the total number of cases, allowing for a flexible range for the total number of infections. Note that the cumulative number of cases ($N_I(t)$) is different from the total number of active cases ($N_A(t)$), representing those currently infected: $N_A^{upp}(t) = A(t) + M(t) + S(t)$ and $N_A^{low}(t) = N_A^{upp}(t) - A(t)$. Cumulative cases grow as a function of time and eventually reach a plateau. In contrast, active cases increase as a function of time reaches a maximum and then starts to fall, eventually reaching zero as A(t), M(t), and S(t) either recover or die. Both cumulative and active cases are given by the model and the user may choose which one to use depending on the data format available for the official cases. In this work, we chose to model our data using the cumulative cases model curves.

To fit the models to the official data we allow the probabilities for asymptomatic and mild/moderate cases to vary in the range $P_A = 0.67 - 0.70$ and $P_M = 0.28 - 0.31$ respectively, while fixing $P_S = 0.02$, according to the medical literature [37, 39, 40]. In practice we let the official curves for cases and deaths tell us exactly which values to use. With recovery times and probabilities for each case severity fixed, we are left with P_D and S to be adjusted. We follow the available information in the literature but ultimately let the models adjust according to the data. We adjust the model parameters and probabilities to get a good fit on the overall shape of the curve (setting S is critical) and determine the best fit by minimizing the χ^2 to obtain a good fit by comparing $D(t)$ to the official deaths (D_O) and verifying which isolation model corresponds to the official data. We then use the same isolation to adjust both $N_I^{upp}(t)$ and $N_I^{low}(t)$ to the official data. Since both cumulative models are well above the official cases we multiply the official case curves by an underestimation correction, meaning that many cases are 'missing' in the official data. We end up with two underestimation factors (U_C^{low} and U_C^{upp}) that give us a range for the number of *real* cases. After both curves are adjusted according to $D(t)$, we have two equations relating the models ($N_I(t)$) and the official cases (C_O): $N_I^{upp}(t) = C_O \times U_C^{upp}$ and $N_I^{low}(t) = C_O \times U_C^{low}$.

To calculate the case fatality rate (CFR) we simply divide official deaths (D_O) by the total number of official cases (C_O), according to equation 2(a). However the model offers a way to correct the denominator of this equation and we use both U_C^{low} and U_C^{upp} to obtain the corrected case fatality rates, according to equations 2(b),(c). These equations properly correct the case fatality rates by taking into account the adjusted number of cases. Note that using U_C^{low} as a correction produces an upper limit for the case fatality rate ($CCRF^{upp}$) and using U_C^{upp} produces a lower limit ($CCFR^{low}$), because $U_C^{low} < U_C^{upp}$.

$$\begin{gathered} CFR = D_O/C_O, (a) \\ CCFR^{low} = D_O/N_I^{upp}(t) = D_O/(C_O \times U_C^{upp}) = CFR/U_C^{upp}, (b) \\ CCFR^{upp} = D_O/N_I^{low}(t) = D_O/(C_O \times U_C^{low}) = CFR/U_C^{low}, (c) \end{gathered} \tag{2}$$

As shown in the next section these corrections dramatically change the case fatality rates for the COVID-19 pandemic.

3. Results

In this section we show: i) The parameter space of the model; ii) The fits of official data for selected cities and countries around the globe; iii) The case fatality rates and their corrections using the model results.

3.1. Exploring the Model Parameter Space

To test the model parameter space we simulated a fictitious city with 1,000,000 individuals using as input (figure 2): $C_F = 0.40$; $\delta = 0.10$; $P_A =$

0.70; $P_M = 0.29$; $P_S = 0.01$; $P_D = 0.4$; $t_A = 10$; $t_M = 15$; $t_S = 20$. On the left side we fix $I = 0\%$ and sweep several values of the infection probability parameter from $S = 10^{-2}$ to $S = 10^{-7}$. If $S \geq 10^{-4}$ we have shorter outbreaks with death curves stabilizing in ~ 200 days (in this particular case) with very high N_I^{upp}, N_I^{low} and severe cases in a very short period of time. Instead, if $S \leq 10^{-5}$ we have longer outbreaks with death curves stabilizing after ~ 600 days. In these cases, N_I^{upp} and N_I^{low}, and severe cases are more spread out in time. On the right side of this figure, we fix $S = 10^{-5}$ and sweep several values of the isolation parameter ranging from $I = 0$ to $I = 0.90$. The most ubiquitous features when I varies are: (i) I changes the height of the plateau reached in each case; ii) different values of I do not impact the duration of the outbreak; iii) The number of cumulative and severe cases change drastically from $I = 0$ to $I = 0.90$ since the numbers of exposed individuals change with I. The number of deaths also varies from 150 ($I = 0.90$) to 2800 ($I = 0$). Model results for the severe cases for different values of S and I have a peak and eventually vanish as more people either die (with probability P_D) or recover (with probability $1 - P_D$). This experiment clearly demonstrates the model's ability to cover different scenarios, ranging from short-lived outbreaks to long-term infections. Both parameters S and I are able to account for changes in the steepness, plateau, duration of outbreaks, and most importantly how rapidly the disease is spread among the population.

3.2. Fitting the probabilistic model to official COVID-19 cases

We present the applicability of the model to different population sizes ranging from cities with ~ 1 million people to larger cities and countries with several million individuals. In every figure of this section we use $C_F = 0.40$, $\delta = 0.10$, $t_A = 10$, $t_M = 15$ and $t_S = 20$ to generate models from $I = 0.30$ to $I = 0.90$. Other parameters and probabilities are adjusted according to each case and are shown in the figures.

Figure 3, left side, displays the models for the city of São Paulo, Brazil (pop. 12,252,000) which accumulated 247,730 official cases and 11,030 deaths in the first 181 days after the first officially reported case on February 26^{th}, 2020. On the right side, we show the results for New York City, USA (pop. 8,336,817) which reached 247,613 cases and 19,196 deaths 228 days after the first reported case on February 29^{th}, 2020. The official data for cases and deaths (red continuous curves) for both cities are from official government sources [41, 42]. Figures 3(a),(b),(c),(d) show the fits for the upper and lower limit models to the official cases for each city. All fits were adjusted multiplying the official data (C_O) by underestimation factors U_C^{upp} and U_C^{low} (indicated in the plots) thus correcting for the missing/undiagnosed cases, as discussed in section 2. The main result is that São Paulo has its official cases underestimated by factors ranging from $U_C^{low} = 4.39$ (lower limit) to $U_C^{upp} = 15.69$ (upper limit). In other words, São Paulo had, according to the model, at least 4.39 times more cases than officially reported. For New York we find underestimation factors ranging from $U_C^{low} = 6.12$ to $U_C^{upp} = 21.87$, meaning that New York had at least 6.12 times more cases than officially reported. Figures 3(e),(f) show the evolution of severe cases for both cities. We find that for São Paulo the severe cases reach a maximum of 8,291 individuals 106 days after the first reported case ($I = 0.70$ model curve). For New York, we find that severe cases reach a maximum of 41,291 such cases 32 days after the first reported case ($I = 0.50$ model curve). The official data for severe cases was not available and is thus not shown. The time evolution of deaths is shown in figures 3(g),(h). We use the official deaths fit corresponding to $I = 0.70$ (São Paulo) and $I = 0.50$ (New York) to adjust both the upper and lower limit models to the official cases.

Another example is shown in figure 4 for Italy (pop. 60,360,000) and Spain (pop. 47,431,256). The official data for these countries were obtained from official government sources [43, 44]. Italy reached 298,200 cases and 35,710 deaths 219 days after the first reported case on February 15^{th}, 2020. The time evolution for the upper and lower limit cases for Italy is shown in figures 4(a),(c). We find $U_C^{low} = 10.13$ and $U_C^{upp} = 33.79$, meaning that Italy had at least 10.13 times more cases than officially reported. Figure 4(e) shows that severe cases in Italy reached a maximum of 129,481 individuals 41 days after the first reported case ($I = 0.70$ curve). The time evolution of deaths in Italy is shown in figure 4(g). Note that we use the official deaths best fit corresponding to a line above the $I = 0.90$ model curve to adjust both the upper and lower limit models shown in figures 4(a),(b). We further assume that this line also fits the severe cases in figure 4(e). According to official data [44], Spain reached 543,400 cases and 29,630 deaths 208 days after the first reported case on February 15^{th}, 2020. The time evolution for both upper and lower limit cases are shown in figures 4(b),(d) and we find $U_C^{low} = 9.22$ and $U_C^{upp} = 29.77$, meaning that Spain had at least 9.22 times more cases than officially reported. Figure 4(f) shows that severe cases in Spain reached a maximum of 50,760 such cases 40 days after the first reported case ($I = 0.70$ curve). The time evolution of deaths is shown in figure 4(h) and a model curve between $I = 0.70$ and $I = 0.90$ is the best fit, which was also used to adjust both upper and lower limit models in figures 4(b),(d). The same approach described above was applied to other cities and countries and the results are summarized in table 1.

3.3. Correcting the Case Fatality Rates (CCFR)

One of the most important and elusive features of the COVID-19 pandemic is the fatality rate since the official cases are highly underestimated [45, 46]. Thus calculating the case fatality rates according to equation 2(a) necessarily produces highly overestimated values (see table 1, CFR(%) column). As described in section 2 the corrected fatality rates are obtained using equations 2(b),(c) and the lower and upper values of U_C obtained from the fits. We calculated the CFR's and their corrections for several countries and cities as shown in table 1. We find mean corrected case fatality rates and standard deviations of $CCFR^{low} = 0.33 \pm 0.02$ and $CCFR^{upp} = 1.14 \pm 0.07$.

4. Discussion

The Parameter Space. We tested the model's ability to properly cover its parameter space and to account for different scenarios, ranging from short-lived outbreaks to long-duration infections. We did this by varying parameters S and I. The parameter S is able to describe several outbreak scenarios as shown in figure 2 (left side) where we fix the value of I to 0 and run simulations for different values of S in the range $1.0 \times 10^{-7} - 1.0 \times 10^{-2}$. In practice the value of S governs the length of the outbreak. The parameter I measures how the fraction of isolated individuals affects the evolution of the pandemic episode. This is investigated in figure 2 (right side) where we plot the time evolution of N_I^{upp}, N_I^{low}, severe cases, and deaths for several values of I ($S = 10^{-5}$). Although the model is able to account for different scenarios, we point that the infection probability curve may not be a *bonafide* representation of reality. Nevertheless, it is a hypothesis that according to our tests correctly describes the time evolution of cases. We also point out that the isolation factor should not be literally interpreted since the fraction of isolated individuals in a given population varies every day. Also, a city's population is not constant and there is a fair amount of people that move in and out of big urban centers on a daily basis. In other words, there is a social dynamic that is important but is not taken into account in our model.

All these factors compete for an oscillating isolation factor and thus limit its meaning and reach.

Fitting real cases. In figures 3 and 4 we show the fits to real data for two world metropolises with a few million inhabitants and two medium-sized countries with several million individuals. In these figures and for the other cities and countries considered we used a range of values for the S parameter ranging from 9×10^{-5} to 1×10^{-2}. The death probabilities ranged from 14% to 22%. We clearly note that good fits were found regardless of the size of the population, with changes in the parameter S and relatively small variations in the probabilities. However, the parameter S is *not* an assumption. Instead, the shape of official deaths and case curves tell us which value of S better describes the data. We clearly see that the overall shape of the curves and the time to reach a plateau were properly recovered by the fits, suggesting similar and universal characteristics of the disease in different locations around the globe and for different population sizes, despite the complex social dynamics and limitations described above.

The immediate gain from each fit is the U_C^{upp} and U_C^{low} values. These numbers are *very* different for each one of the places considered and are summarized in table 1. For New York, official data indicates 247,730 cases (after 228 days), but we find $U_C^{upp} = 21.87$ and $U_C^{low} = 6.12$, meaning that the *real* number of cases could be at least 6.12 times higher or even 21.87 times higher. If we consider the lower limit as true, for instance, we find that official cases are roughly 16% of the *real* cases, i.e., 84% of the cases went unnoticed. Italy had 298,200 official cases (after 219 days) and we find $U_C^{upp} = 33.79$ and $U_C^{low} = 10.13$. Again, using the lower limit result, we find that Italy had at least 10.13 times more cases than officially reported, meaning that official cases represent roughly 10% of the *real* cases. Similar conclusions might be drawn for the other places considered and this discrepancy in official and *real* cases could be due to limited or no access to medical facilities, poor testing or a myriad of other factors [47, 48].

Severe Cases and ICU needs. Our model estimates the number of severe cases, those that will most likely visit the hospital in need of medical care. Some of these cases might require long-term medical attention and a fraction of those will need intensive care units (ICU) and respirators. The curves corresponding to severe cases show that they go up, reach a peak, and start to decrease as the outbreak progresses. For Spain (see Fig.4), considering the $I = 0.70$ model curve, we see that severe cases peaked roughly at 50,000 individuals. This result alone dramatically explains why the medical facilities and hospitals in Spain (and in many other cities and countries) had exhausted their capacities very early into the outbreak [49, 50, 51, 52]. For instance, if only 2% of these severe cases require intensive care, we are still talking about 1000 ICU beds that must be dedicated to COVID-19 patients. However, as of 2013, Spain had a total of 4700 multi-purpose ICU beds in the entire country [53]. Another example is New York City which had, according to official data 17,259 individuals hospitalized 32 days after the first reported case on February 29^{th}, 2020 [41]. No data was available to show how many of those hospitalized required respirators/ICU beds. For NYC, the model predicts a maximum of 41,291 severe cases for the 32^{nd} day after the first reported case. Taking into account that severe cases take ~ 20 days to recover (or evolve to death) [36] we should compare the model prediction to day 52 after the first reported case. Comparing the model prediction to the official data [41] on day 52 we indeed see that NYC had 41,516 hospitalizations. This demonstrates how precisely the model estimates the number of those in need of hospitalizations/ICU beds for such a full-blown pandemic episode. However, we did not explore in depth this particular aspect of the model since we focused on estimating the corrections to the case fatality rates.

Correcting the Official Case Fatality Rate. We use the U_C^{low} and U_C^{upp} values to correct the official case fatality rates (CFR's) according to equation 2. These results are in table 1. We notice the values of U_C^{low} and U_C^{upp} are *very* different for each of the locations considered but we find $CCFR^{upp} = 1.14\% \pm 0.07\%$ and $CCFR^{low} = 0.33\% \pm 0.02\%$. These small standard deviations for the mean CCFR's contrast to the wildly different CFR's for the many locations considered which average $6.75\% \pm 3.11\%$. Although the corrections applied to the models to fit the many official data curves were very different, they translate into consistent corrected case fatality rates (see table 1). In summary, considering the ubiquitous unreported cases the resulting corrected case fatality rates (CCFR's) are much lower than the case fatality rates (CFR's).

In summary we find the corrected case fatality rates (CCFR's) ranging from 0.33% to 1.14%. This is in accordance with completely independent studies that use sample testing conducted worldwide and especially in hard-hit countries like Brazil and the USA. In particular, a study compiling data from several independent seroprevalence studies in 51 different countries finds case fatality rates varying from 0% to 1.54% with a median of 0.27%. This study includes both light (e.g. Afghanistan) and hard (e.g. USA) hit areas around the globe [54]. Another seroprevalence study conducted in Brazil in

90 cities (comprising 54 million individuals) finds that the official number of cases is underestimated by a factor of 7. These cities combined had (until May 2020) $\sim 105,000$ official cases and $\sim 8,000$ deaths, suggesting a corrected case fatality rate of $\sim 1.10\%$ [55, 56]. Finally, another seroprevalence study conducted in New York City finds a corrected case fatality rate for the first wave of the COVID-19 pandemic of 0.97% [57]. These studies find corrected case fatality rates well inside the range indicated by our model. We highlight that seroprevalence studies are a completely different approach compared to our mathematical-statistical study.

We mention that calculating the *real* case fatality rates for the COVID-19 pandemic is a challenge since it is impacted by access to medical facilities, social constraints, population size and density, the mean age of the impacted population, and a myriad of other factors. That is why our method tries to set both an upper and a lower limit for the CCFR's. We argue that countries with better infrastructure, medical facilities, and social standards might experience a lower case fatality rate than countries with less favorable conditions. Nevertheless, even considering the lower limit of $0.33\% \pm 0.02\%$ this disease is at least 30 times deadlier than the 2009 Swine Flu pandemic which had an estimated case fatality rate of 0.01% [58, 59].

5. Conclusions

Our probabilistic method was able to generate model curves and good fits to the official data for different population sizes, thus demonstrating that the parameters and probabilities chosen to describe the problem for each case are correctly describing the time evolution of the COVID-19 pandemic. These results clearly show how isolation plays an important role in preventing the spread of this disease. This probabilistic model was designed to estimate the number of undiagnosed patients using both an upper and a lower limit for the number of cases, clearly showing that *real* cases might be several times larger than official reported cases, independent of population size and other factors. This clearly demonstrates a universal characteristic for the disease and especially for the corrected case fatality rates. The model also demonstrated that we can estimate the time evolution of severe cases and showed that thousands of patients were in this condition during the most severe days of the outbreak, in accordance with official data. That was the reason for so many deaths in the first months after the first reported cases: hospitals were *not* prepared for a fast-spreading disease like COVID-19. No country or health system can deal with several thousand individuals in need of medical care at the same time, some of which in dire need of respirators. Our main result is the use of a range of values for the number of *real* cases to correct the official fatality rates. Notwithstanding, the corrections applied for the number of cases were very different the derived corrected fatality rates are rather consistent and have small standard deviations which agree with other studies employing seroprevalence methods. We find that the COVID-19 pandemic may not be as deadly as initially thought or as indicated by the CFR's since the corrected fatality rates range from $\mathbf{0.33\% \pm 0.02\%}$ to $\mathbf{1.14\% \pm 0.07\%}$. This revealed that although the pandemic affected countries and regions in different ways we were still able to find a universal characteristic for this pandemic.

References

[1] H. A. Rothan, S. N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak, Journal of autoimmunity (2020) 102433.

[2] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, R. Siddique, Covid-19 infection: Origin, transmission, and characteristics of human coronaviruses, Journal of Advanced Research (2020).

[3] Z. Shi, Z. Hu, A review of studies on animal reservoirs of the sars coronavirus, Virus research 133 (2008) 74–87.

[4] H. Lu, C. W. Stratton, Y.-W. Tang, Outbreak of pneumonia of unknown etiology in wuhan, china: The mystery and the miracle, Journal of medical virology 92 (2020) 401–402.

[5] The official web page for the world health organization (who), 2020. URL: `https://www.who.org`.

[6] I. I. Bogoch, A. Watts, A. Thomas-Bachli, C. Huber, M. U. Kraemer, K. Khan, Pneumonia of unknown aetiology in wuhan, china: potential for international spread via commercial air travel, Journal of travel medicine 27 (2020) taaa008.

[7] S. Zhao, Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, et al., Preliminary estimation of the basic reproduction

number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak, International journal of infectious diseases 92 (2020) 214–217.

[8] Z. Xu, L. Shi, Y. Wang, J. Zhang, L. Huang, C. Zhang, S. Liu, P. Zhao, H. Liu, L. Zhu, et al., Pathological findings of covid-19 associated with acute respiratory distress syndrome, The Lancet respiratory medicine 8 (2020) 420–422.

[9] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in wuhan, china, The lancet 395 (2020) 497–506.

[10] L. Morawska, J. W. Tang, W. Bahnfleth, P. M. Bluyssen, A. Boerstra, G. Buonanno, J. Cao, S. Dancer, A. Floto, F. Franchimon, et al., How can airborne transmission of covid-19 indoors be minimised?, Environment international 142 (2020) 105832.

[11] D. Banerjee, M. Rai, Social isolation in covid-19: The impact of loneliness, 2020.

[12] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun, et al., Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts, The Lancet Global Health (2020).

[13] A. Sharma, R. Fölster-Holst, M. Kassir, J. Szepietowski, M. Jafferany, T. Lotti, M. Goldust, The effect of quarantine and isolation for covid-19 in general population and dermatologic treatments, Dermatol Ther 10 (2020) e13398.

[14] F. Durankuş, E. Aksu, Effects of the covid-19 pandemic on anxiety and depressive symptoms in pregnant women: a preliminary study, The Journal of Maternal-Fetal & Neonatal Medicine (2020) 1–7.

[15] M. Nicola, Z. Alsafi, C. Sohrabi, A. Kerwan, A. Al-Jabir, C. Iosifidis, M. Agha, R. Agha, The socio-economic implications of the coronavirus pandemic (covid-19): A review, International journal of surgery (London, England) 78 (2020) 185.

[16] Corona virus resource center, 2020.

[17] H. W. Hethcote, The mathematics of infectious diseases, SIAM review 42 (2000) 599–653.

[18] A. W. Roddam, Mathematical epidemiology of infectious diseases: Model building, analysis and interpretation: O diekmann and jap heesterbeek, 2000, chichester: John wiley pp. 303,£ 39.95. isbn 0-471-49241-8, 2001.

[19] L. Peng, W. Yang, D. Zhang, C. Zhuge, L. Hong, Epidemic analysis of covid-19 in china by dynamical modeling, arXiv preprint arXiv:2002.06563 (2020).

[20] N. H. Tuan, H. Mohammadi, S. Rezapour, A mathematical model for covid-19 transmission by using the caputo fractional derivative, Chaos, Solitons & Fractals 140 (2020) 110107.

[21] K. Chatterjee, K. Chatterjee, A. Kumar, S. Shankar, Healthcare impact of covid-19 epidemic in india: A stochastic mathematical model, Medical Journal Armed Forces India (2020).

[22] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, et al., Early dynamics of transmission and control of covid-19: a mathematical modelling study, The lancet infectious diseases (2020).

[23] W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics, Proceedings of the Royal society A, Mathematical, Physical and Engineering Sciences 115 (1927) 700–721.

[24] T. Harko, F. S. N. Lobo, M. K. Mak, Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates, Applied Mathematics and Computation 236 (2014) 184 – 194.

[25] H. W. Hethcoat, The mathematics of infectious diseases, SIAM Rev. 42 (2000) 599–653.

[26] A. Huppert, G. Katriel, Mathematical modelling and prediction in infectious disease epidemiology, Clinical Microbiology and Infection 19 (2013) 999 – 1005. doi:https://doi.org/10.1111/1469-0691.12308.

[27] D. M. Jenkins, An examination of mathematical models for infectious disease, Honors Research Projects 194, University of Akron (2015).

[28] G. Onder, G. Rezza, S. Brusaferro, Case-fatality rate and characteristics of patients dying in relation to covid-19 in italy, Jama 323 (2020) 1775–1776.

[29] D. D. Rajgor, M. H. Lee, S. Archuleta, N. Bagdasarian, S. C. Quek, The many estimates of the covid-19 case fatality rate, The Lancet Infectious Diseases 20 (2020) 776–777.

[30] D. Baud, X. Qi, K. Nielsen-Saines, D. Musso, L. Pomar, G. Favre, Real estimates of mortality following covid-19 infection, The Lancet infectious diseases (2020).

[31] P. Spychalski, A. Błażyńska-Spychalska, J. Kobiela, Estimating case fatality rates of covid-19, The Lancet. Infectious Diseases (2020).

[32] M. Lipsitch, Estimating case fatality rates of covid-19, The Lancet. Infectious Diseases (2020).

[33] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, Numerical Recipes in Fortran. The Art of Scientific Computing, 2nd ed., Press Syndicate of the University of Cambridge, 1992.

[34] J. M. van Seventer, N. S. Hochberg, Principles of infectious diseases: Transmission, diagnosis, prevention, and control, International Encyclopedia of Public Health (2017) 22.

[35] R. T. Gandhi, J. B. Lynch, C. del Rio, Mild or moderate covid-19, New England Journal of Medicine (2020).

[36] H. C. Prescott, T. D. Girard, Recovery from severe covid-19: leveraging the lessons of survival from sepsis, Jama 324 (2020) 739–740.

[37] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S.-m. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, et al., Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19), International journal of infectious diseases 94 (2020) 154.

[38] A. Lachmann, Correcting under-reported covid-19 case numbers, medRxiv (2020).

[39] C. C.-. R. Team, C. C.-. R. Team, C. C.-. R. Team, S. Bialek, E. Boundy, V. Bowen, N. Chow, A. Cohn, N. Dowling, S. Ellington, et al., Severe outcomes among patients with coronavirus disease 2019 (covid-19)—united states, february 12–march 16, 2020, Morbidity and mortality weekly report 69 (2020) 343–346.

[40] Y. Liu, L.-M. Yan, L. Wan, T.-X. Xiang, A. Le, J.-M. Liu, M. Peiris, L. L. Poon, W. Zhang, Viral dynamics in mild and severe cases of covid-19, The Lancet Infectious Diseases (2020).

[41] The official website of the city of new york, https://www1.nyc.gov/. Accessed: October 2020, 2020.

[42] Geocovid portal, ufes, http://covid.mapbiomas.org/. Accessed: September 2020, 2020.

[43] The official website of the italian ministry of health, https://www.worldometers.info/coronavirus/country/italy/. Accessed: September 2020, 2020.

[44] The official website of the spainish ministry of health, https://www.worldometers.info/coronavirus/country/spain. Accessed: September 2020, 2020.

[45] S. L. Wu, A. Mertens, Y. Crider, et al., Substantial underestimation of sars-cov-2 infection in the united states, Nature Commun (2020) 4507.

[46] D. Böhning, I. Rocchetti, A. Maruotti, H. Holling, Estimating the undetected infections in the covid-19 outbreak by harnessing capture–recapture methods, International Journal of Infectious Diseases 97 (2020) 197 – 201. doi:10.1016/j.ijid.2020.06.009.

[47] N. A. Alwan, Surveillance is underestimating the burden of the covid-19 pandemic, The Lancet 396 (2020) e24.

[48] S. Burgess, M. J. Ponsford, D. Gill, Are we underestimating seroprevalence of sars-cov-2?, 2020.

[49] L. Carenzo, E. Costantini, M. Greco, F. Barra, V. Rendiniello, M. Mainetti, R. Bui, A. Zanella, G. Grasselli, M. Lagioia, et al., Hospital surge capacity in a tertiary emergency referral centre during the covid-19 outbreak in italy, Anaesthesia (2020).

[50] C. Massonnaud, J. Roux, P. Crépey, Covid-19: Forecasting short term hospital needs in france, medRxiv (2020).

[51] C. Bernucci, C. Brembilla, P. Veiceschi, Effects of the covid-19 outbreak in northern italy: perspectives from the bergamo neurosurgery department, World neurosurgery 137 (2020) 465.

[52] M. Sorbello, K. El-Boghdadly, I. Di Giacinto, R. Cataldo, C. Esposito, S. Falcetta, G. Merli, G. Cortese, R. Corso, F. Bressan, et al., The italian coronavirus disease 2019 outbreak: recommendations from clinical practice, Anaesthesia 75 (2020) 724–732.

[53] M. Martín, C. León, J. Cuñat, F. del Nogal, Intensive care services resources in spain, Medicina Intensiva 37 (2013) 443–451. doi:`10.1016/j.medine.2013.06.002`.

[54] J. Ioannidis, The infection fatality rate of covid-19 inferred from seroprevalence data, medRxiv (2020).

[55] P. C. Hallal, F. C. Barros, M. F. Silveira, A. J. D. d. Barros, O. A. Dellagostin, L. C. Pellanda, C. J. Struchiner, M. N. Burattini, F. P. Hartwig, A. M. B. Menezes, et al., Epicovid19 protocol: repeated serological surveys on sars-cov-2 antibodies in brazil, Ciencia & saude coletiva 25 (2020) 3573–3578.

[56] P. C. Hallal, et al., Covid-19 no brasil: várias epidemias num sópaísprimeira fase do epicovid19 reforça preocupação com a região norte, https://wp.ufpel.edu.br/covid19/files/2020/05/EPICOVID19BR-release-fase-1-Portugues.pdf. Accessed: January 2021, 2020.

[57] D. Stadlbauer, J. Tan, K. Jiang, M. M. Hernandez, S. Fabre, F. Amanat, C. Teo, G. A. Arunkumar, M. McMahon, C. Capuano, et al., Repeated cross-sectional sero-monitoring of sars-cov-2 in new york city, Nature (2020) 1–5.

[58] S. Riley, K. O. Kwok, K. M. Wu, D. Y. Ning, B. J. Cowling, J. T. Wu, L.-M. Ho, T. Tsang, S.-V. Lo, D. K. Chu, et al., Epidemiological characteristics of 2009 (h1n1) pandemic influenza based on paired sera from a longitudinal community cohort study, PLoS Med 8 (2011) e1000442.

[59] J. Y. Wong, D. K. Heath Kelly, J. T. Wu, G. M. Leung, B. J. Cowling, Case fatality risk of influenza a (h1n1pdm09): a systematic review, Epidemiology (Cambridge, Mass.) 24 (2013).

Acknowledgements

We would like to thank the Federal University of Rio Grande do Norte (UFRN), especially the School of Science and Technology (ECT).

Author contributions statement

J.R.S.L conceived the model and conducted the simulations, G.Z.S.L analysed the results, J.R.S.L. prepared figures, G.Z.S.L revised the code. Both authors wrote and reviewed the manuscript.

Additional information

Competing interests The authors declare no competing interests.

Place (pop. [10^6])	**Cases** [10^3]	**Deaths** [10^3]	U_C^{low}	U_C^{upp}	$CFR(\%)$	$CCFR^{upp}(\%)$	$CCFR^{low}(\%)$
Natal, Brazil (0.88)	26.89	1.09	3.37	12.00	4.05	1.20	0.34
Curitiba, Brazil (1.93)	40.91	1.42	3.15	10.82	3.47	1.10	0.32
Fortaleza, Brazil (2.64)	49.21	3.87	6.61	23.69	7.86	1.19	0.33
Manaus, Brazil (2.78)	50.02	2.51	4.71	15.68	5.02	1.07	0.32
Rio de Janeiro, Brazil (6.72)	151.89	14.02	7.62	27.23	9.23	1.21	0.34
New York, USA (8.34)	247.61	19.20	6.12	21.87	7.75	1.27	0.35
Mexico City, Mexico (8.92)	107.60	8.85	7.63	27.21	8.22	1.08	0.30
São Paulo, Brazil (12.25)	247.73	11.03	4.39	15.69	4.45	1.01	0.28
Spain* (47.43)	281.00	28.40	9.22	29.77	10.11	1.10	0.34
South Korea* (51.64)	14.63	0.31	1.92	6.88	2.12	1.10	0.31
Italy* (60.36)	298.20	35.71	10.13	33.79	11.98	1.18	0.35
					$\mathbf{6.75 \pm 3.11}$	$\mathbf{1.14 \pm 0.07}$	$\mathbf{0.33 \pm 0.02}$

Table 1: Summary of results for several cities and countries (marked with*) with different populations. We use the number of cases and deaths to calculate the case fatality rates. The model fits yield U_C^{low} and U_C^{upp} which are then used to calculate both $CCFR^{upp}$ and $CCFR^{low}$. Note that although CFR's vary for each location, the averaged $CCFR^{upp}$ and $CCFR^{low}$ standard deviations are comparatively smaller.

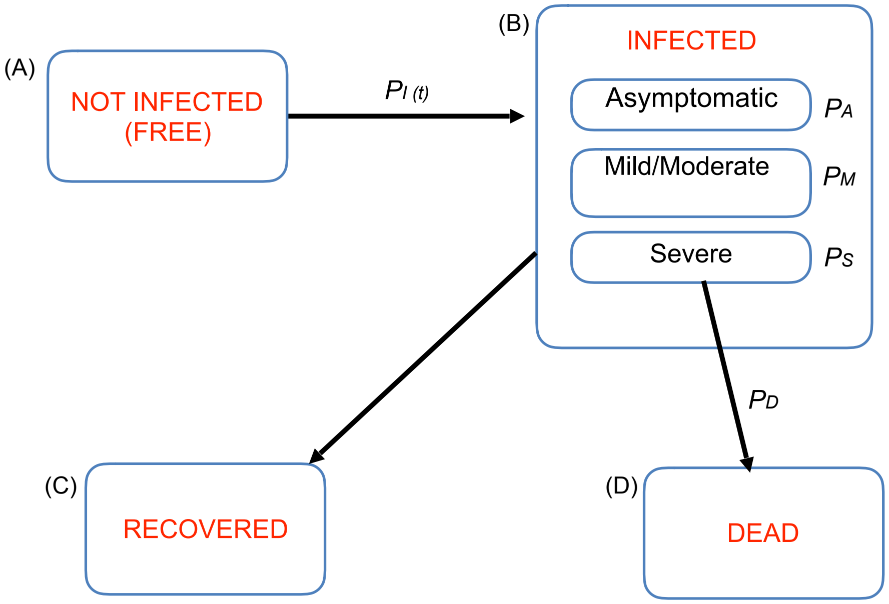

Figure 1: **The schematic diagram of the probabilistic model of COVID-19 cases**. Each individual might be in one of these compartments: free (A), infected with 3 subclasses (B), recovered (C), or dead (D). Once infected, there are three possible conditions that precede recovery or death: asymptomatic (P_A), mild/moderate (P_M) and severe (P_S) [B]. Eventually, Asymptomatic and mild/moderate individuals always recover after times t_A and t_M respectively (C). After a time t_S, severe cases either recover (C), or die (D), with the probabilities $P_R = 1 - P_D$ and P_D respectively.

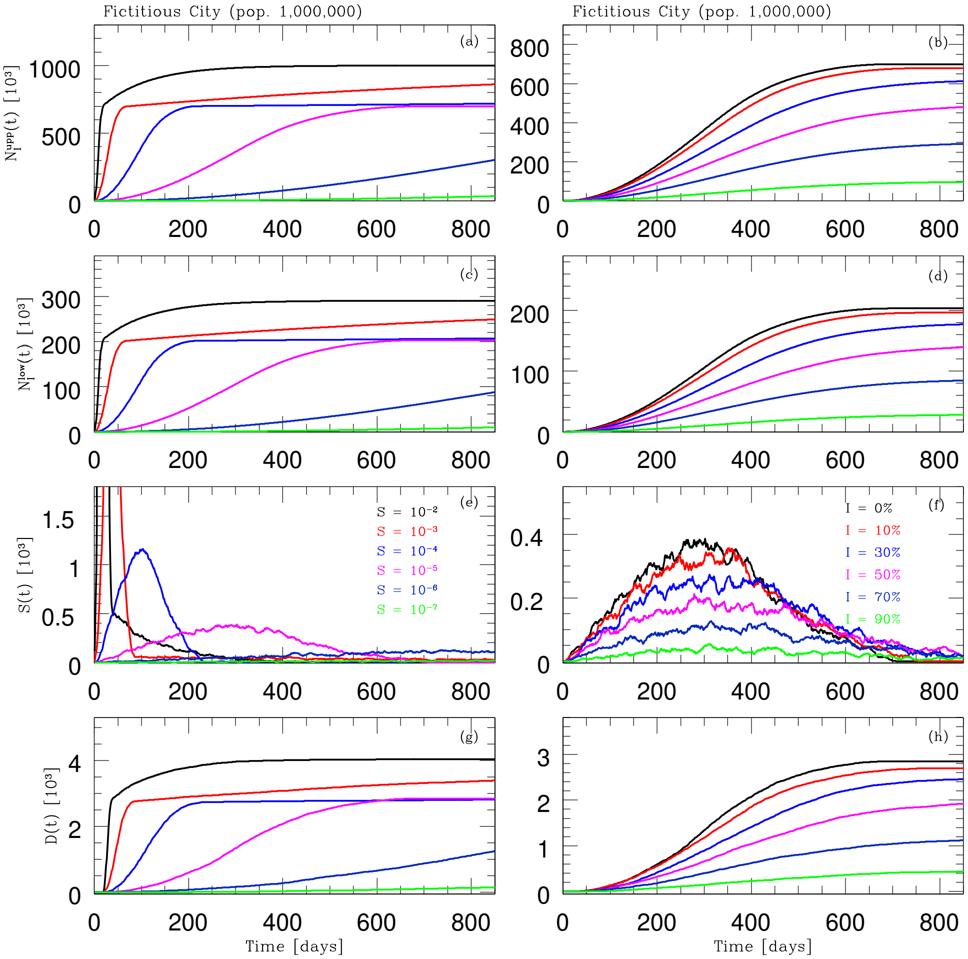

Figure 2: **Exploring the model parameter space.** This figure shows a variety of scenarios for the evolution of the COVID-19 disease for different parameters. We simulate a fictitious city with 1,000,000 inhabitants for a variety of values of the isolation factor (I) and for several values of the probability curve $P_I(t)$ parameter S (equation 1). On the left hand side panels we show various scenarios referring to the different values of the parameter S (with a fixed isolation factor $I = 0\%$). Each panel on the left-hand side shows a different result: The upper limit for the total number of cases (a), the lower limit for the total number of cases (c), the severe cases (e) and deaths (g). On the right hand side panels we show several scenarios referring to the different values of I (with fixed S = 10^{-5}). Panels (b), (d), (f) and (h) also show the upper limit for the total number of cases, the lower limit for the total number of cases, severe cases and deaths, respectively.

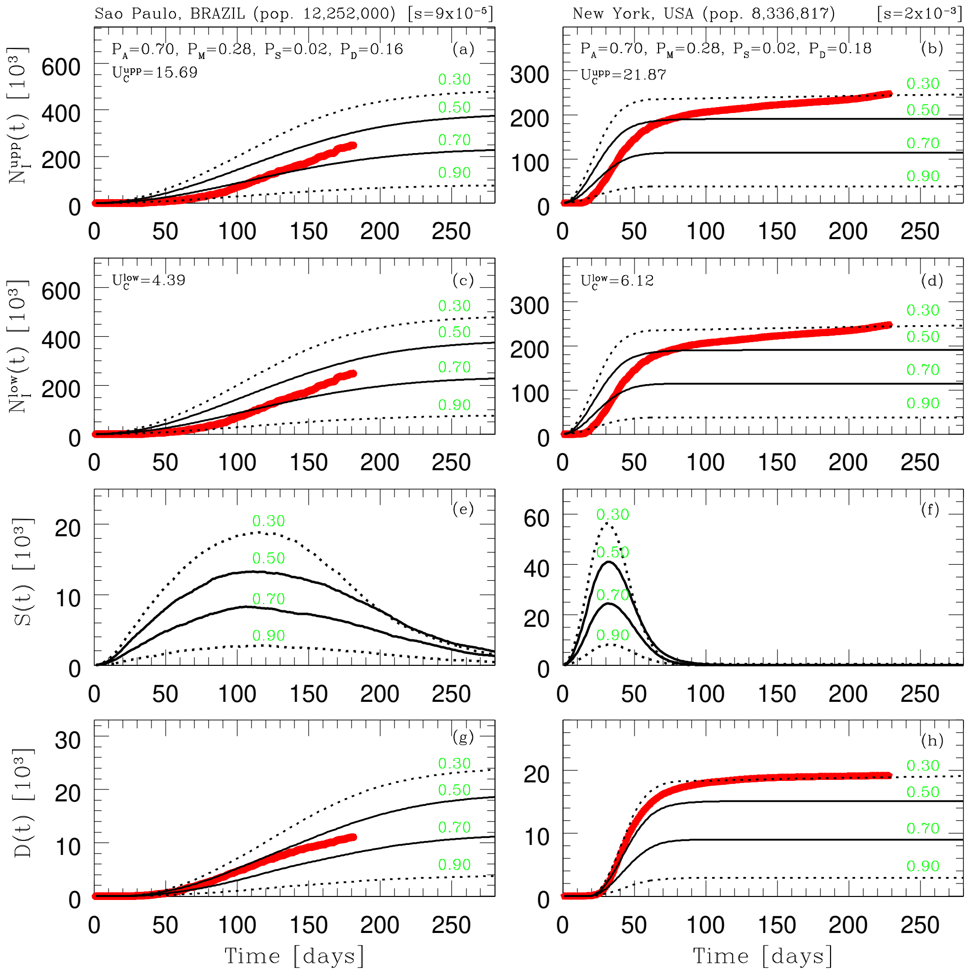

Figure 3: **Probabilistic model fitting results to real COVID-19 cases for two global megalopolis.** In this figure we show the models (continuous and dashed black curves) adjusted for the real data (continuous red curves) for two megalopolis: São Paulo -SP (left-hand side) and New York - NY (right-hand side). Each model curve corresponds to a different isolation factor I (green), covering most of the parameter space from $I = 30\%$ to $I = 90\%$. The probabilities P_A, P_M, P_S and P_D and the parameters S and U_C used to adjust the models to the data are indicated. We show the upper limit for the total number of cases (a) and (b); the lower limit for the total number of cases (c) and (d); severe cases (e) and (f) and deaths cases (g) and (h). Our model estimates the severe cases in each considered scenario, although we do not have the time evolution of real severe cases for each of the cities considered.

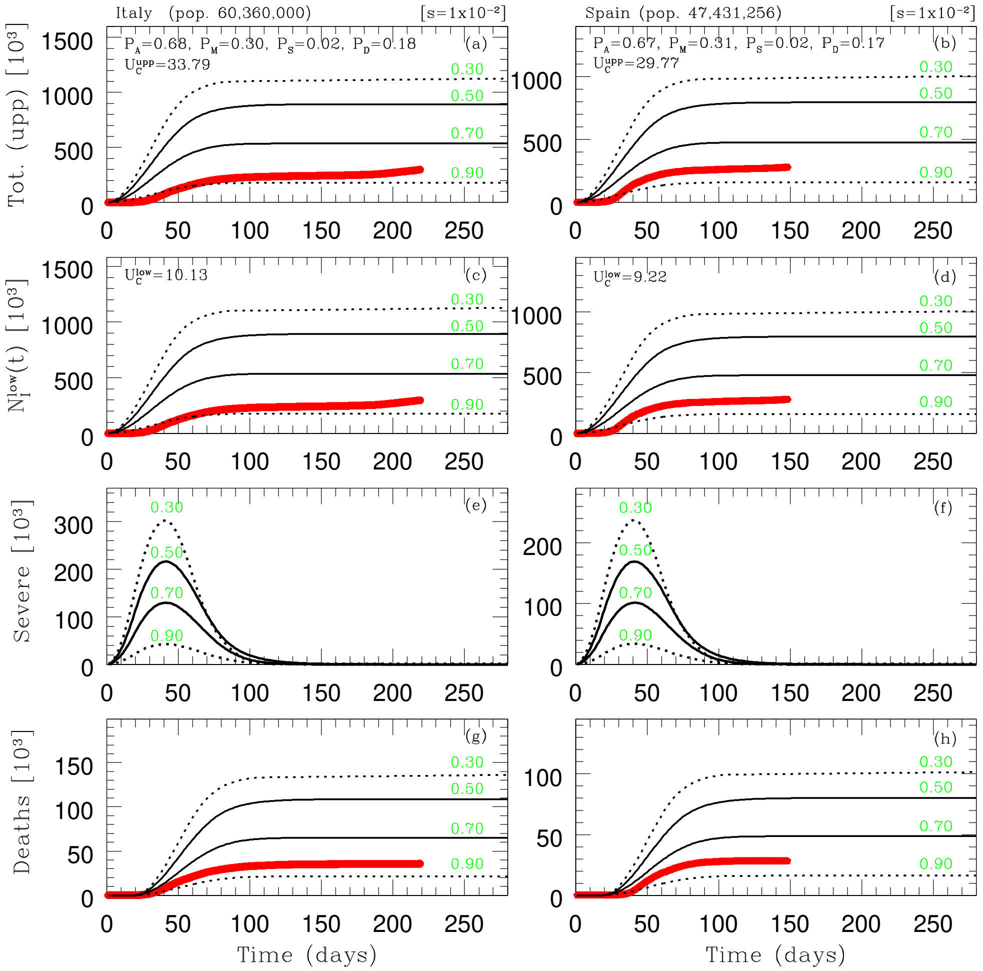

Figure 4: **Probabilistic model fitting results to real COVID-19 cases for two European countries** In this figure we show the models (continuous and dashed black curves) adjusted for the real data (continuous red curves) for two European countries: Italy (left-hand side) and Spain (right-hand side). Each model curve corresponds to a different isolation factor I (green), covering most of the parameter space from $I = 30\%$ to $I = 90\%$. The probabilities P_A, P_M, P_S and P_D and the parameters S and U_C used to adjust the models to the data are indicated. We show the upper limit for the total number of cases (a) and (b); the lower limit for the total number of cases (c) and (d); severe cases (e) and (f) and deaths cases (g) and (h). Although we do not have the time evolution of real severe cases for each of the cities considered, our model estimates the severe cases in each investigated scenario.