From a single host to global spread: The global mobility based modelling of the COVID-19 pandemic implies higher infection rate and lower detection ratio than current estimates.

Marian Siwiak PhD 1, Pawel Szczesny PhD 1, Marlena Siwiak PhD 1
1 Data 3.0, Sevenoaks, TN14 7TG, UK
Corresponding Author:
Marian P. Siwiak PhD 1
3 Forge Way, Sevenoaks, TN14 7TG, United Kingdom
Email address: siwiak@data30.co.uk

Abstract

Background

Since the outbreak of the COVID-19 pandemic, multiple efforts of modelling of the geo-temporal transmissibility of the virus have been undertaken, but none describes the pandemic spread at the global level. The aim of this research is to provide a high-resolution global model of the pandemic that overcomes the problem of biased country-level data on the number of infected cases. To achieve this we propose a novel SIR-type metapopulation transmission model and a set of analytically derived model parameters. We used them to perform a simulation of the disease spread with help of the Global Epidemic and Mobility (GLEAM) framework embedding actual population densities, commute patterns and long-range travel networks. The simulation starts on Nov 17th, 2019 with just a single pre-symptomatic, yet infectious, case in Wuhan, China, and results in an accurate prediction of the number of diagnosed cases after 154 days in multiple countries across five continents. In addition, the model outcome shows high compliance with the results of a random screening test conducted on pregnant women in the New York area.

Methods

We have built a modified SIR metapopulation transmission model and parameterized it analytically either by setting the values of the parameters based on the literature, or by assuming their plausible values. We compared our results with the number of diagnosed cases in twenty selected countries which provide reliable statistics but differ substantially in terms of strength and speed of undertaken Non-Drug Interventions. The obtained 95% confidence intervals for the predictions are in agreement with the empirical data.

Results

The parameters that successfully model the pandemic are: the basic reproduction number $\boldsymbol{R_0}$, 4.4; a latent non-infectious period of 1.1. days followed by 4.6 days of the presymptomatic infectious period; the probability of developing severe symptoms, 0.01; the probability of being diagnosed when presenting severe symptoms of 0.6; the probability of diagnosis for cases with mild symptoms or asymptomatic, 0.001.

Discussion

Parameters that successfully reproduce the observed number of cases indicate that both R_0 and the prevalence of the virus might be underestimated. This is in concordance with the newest research on undocumented COVID-19 cases. Consequently, the actual mortality rate is putatively lower than estimated. Confirmation of the pandemic characteristic by further refinement of the model and screening tests is crucial for developing an effective strategy for the global epidemiological crisis.

Introduction

As of April 23rd, 2020, novel coronavirus SARS-CoV-2 has already spread into 185 countries and territories around the world (Dong, Du & Gardner, 2020). With over two and half a million confirmed infections and over 180 thousand deaths (Dong, Du & Gardner, 2020), it became a global challenge. COVID-19, the disease caused by this coronavirus, was characterised as a pandemic by WHO on March 11th, 2020.

While several different measures to contain the virus have been implemented by countries all over the world, their effectiveness remains to be seen. Until an effective treatment is available, the accuracy of the pandemic models and the decisions made on their basis are the major factors in reducing the overall mortality in the COVID-19 pandemic.

The models used to inform decision-makers differ significantly in their basic assumptions because it is the first coronavirus of such an impact in terms of the number of fatal cases. Moreover, the existing modelling approaches often use biased data for tuning parameters or assessing models' quality. In particular, most models use country-level data which is biased by one or many of the following factors: (i) level of transparency in acquiring, analysing, interpreting and reporting of data, (ii) level of detection effort, (iii) efficiency of introduced Non-Drug Interventions, (iv) biased sampling of people to be tested (individuals showing severe and typical symptoms or suspected of having a contact with an infected person are more likely to be tested). The bias in data could be avoided by conducting well designed, random screening tests, but so far just a few such attempts have been made and they were limited to small and isolated communities in Italy (Lavezzo et al., 2020), and Germany (Streeck et al., 2020), or women admitted for delivery (Sutton et al., 2020).

Also, as of the date of submitting this work, there were no peer-reviewed geo-temporal models of the pandemic. We argue that creating a global model by fitting curves to observed data is impossible, unless the data used in this exercise is a result of large scale screening tests. As mentioned above, country-level data is biased and the nature of this bias is different for each country, mainly reflecting state policy towards the disease. Obviously, fitting model curves on global data bears a significant error, as this data is a mixture of all country-level biases.

Even the proper characteristic of the virus is hampered by the above-mentioned biases. For instance, early estimates of the basic reproduction number of the virus, R_0, were typically obtained using only Chinese data on the number of diagnosed cases (Zhang et al., 2020; Wu, Leung & Leung, 2020; Liu et al., 2020). These estimates proposed the value of R_0 within the range 1.5-6.47, and the earliest most likely served as the basis for the January's official WHO estimate, which stated R_0=1.4-2.5 ("Statement on the meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus 2019 (n-CoV) on 23 January 2020"). However, re-analysis of Chinese data provided an updated estimate of 5.7 (95% CI 3.8-8.9) (Steven Sanche et al., 2020).

Similar problems arise when estimating the actual prevalence of the virus. In this case, the estimates are not only biased by the state policy, but also by the fact that many infections are mild, asymptomatic or with atypical symptoms. In fact, many COVID-19 cases pass unnoticed, for instance in Li's and coworkers China study, over 50% (Li et al., 2020a).

The main objective of this research is to create a global model of the early stages of the pandemic that would overcome the problem of the biased data on the number of infected cases. This goal was achieved by creating a first global model of the COVID-19 pandemic that builds on top of the successful modelling framework GLEAM (Van den Broeck et al., 2011). In contrast to many existing models, our attempt did not use biased country-level data on the number of infected cases to fit the model curve. Instead, it used a set of predefined parameters to simulate the spread of the disease around the world starting from a single infected host in Wuhan, China on Nov 17th, 2019. The exact date was suggested by unverified press reports and used widely as a date of a disease contraction for "patient zero", but evolutionary tracing also suggests a similar timepoint (Li et al., 2020b). The simulation took into account the current population densities all over the world, actual commute and flight networks, and travel ability of infected individuals. The simulation was run with 20 iterations for 154 days till Apr 19th, 2020. The obtained results were used to create 95% confidence intervals for curves of cumulative number of diagnosed cases, separately for each country in the world.

To assess the quality of our model, its results should be compared to the observed data on the number of diagnosed cases, but this data suffers from four biases, listed above. These biases cannot be totally eliminated, but careful selection of countries used in the analysis may reduce their impact significantly. In particular, we limited the analysis of model results to twenty countries which to our belief provide the most accurate and transparent reports on the number of infected cases (reducing bias nr 1). Selected countries are also divergent in terms of their detection efforts illustrated by the number of conducted tests per capita (reducing bias nr 2). Additionally, our model depicts the early stages of the pandemic when Non-Drug Interventions were not yet introduced on a large scale in the selected countries, and if they were, the exact time of intervention was added to the final diagrams to show its impact on the number of diagnosed cases (reducing bias nr 3). Furthermore, the proportion of symptomatic versus mild and asymptomatic cases is built in the model, so is the fact that symptomatic individuals are more likely to be tested (reducing bias nr 4). Finally, the detectability of the disease is also built in the model. This means that the presented confidence intervals depict the plausible range of diagnosed cases assuming a given detectability, not the actual number of infected individuals in the country. However, the latter may easily be assessed knowing the assumed detectability, or derived from the model file provided with this paper.

The presented model enables better understanding of the virus, its infectivity and detectability. Also, it may serve as a solid foundation for further attempts of global and country-level modelling. In particular, more detailed models that include information on introduced precautions may be created by making the detectability parameter variable in time and geographics. This would enable an optimal pandemic strategy to be established for each country.

Materials & Methods

Modelling software

The model is based on the Global Epidemic and Mobility Model (GLEAM) framework (Balcan et al., 2010), implemented in the GLEAMviz software (Van den Broeck et al., 2011). The GLEAM model integrates sociodemographic and population mobility data in a spatially structured stochastic disease approach to simulate the spread of epidemics at a worldwide scale. It was previously used for a real-time numerical forecast of the global spreading of Influenza A/H1N1 (Tizzoni et al., 2012), and the accuracy of that modelling was later confirmed (Tizzoni et al., 2012).

Data sources

The reference data about the number of SARS-CoV-2 diagnosed patients in the period from Jan 22, 2020, to Apr 19th, 2020, were downloaded from the Johns Hopkins University of Medicine Coronavirus Resource Center GitHub repository https://github.com/CSSEGISandData/COVID-19. The provided data have been grouped by country. These data sources were used to assess the quality of the model results. Empirical data for the time preceding Jan 22nd, 2020 is not available in the cited source.

Information about the severity of developed symptoms was derived from the worldometer.info website https://www.worldometers.info/coronavirus/.

Information on testing efforts in selected countries, comes from the https://ourworldindata.org/coronavirus-testing-source-data website.

Approximation of the number of mild and asymptomatic cases in the New York area was derived from the results of a random screening test performed on women admitted for delivery at the New York–Presbyterian Allen Hospital and Columbia University Irving Medical Center (Sutton et al., 2020).

Information on introduced Non-Drug Interventions comes from publicly available sources.

Other data sources, such as subpopulation selection, commuting patterns, or air travel flows, used during simulation, are embedded in the GLEAM software and well described by its developers (Van den Broeck et al., 2011).

Model parametrization

Below and in **Table 1,** we present a set of parameters that was used in the model. Most parameters were derived from the literature. In the absence of a reliable reference, the parameters were assigned with the most plausible values by the authors based on the epidemiological knowledge on SARS-CoV-2 and other viruses. The parameters' derivation method is summarized in **Table 1**.

The average latency period (***lp***) of 5.6 days is a consensus of different estimations calculated previously (Lauer et al., 2020). Due to 1) long ***lp***, effectively much longer than reported for other coronaviruses, and 2) known cases of presymptomatic transmission (Woelfel et al., 2020; Tong et al., 2020), for the modelling purposes we decided to split the latency period into two parts: 1) average latent non-infectious period (***lnip***) of *1.1* days (based on the time of infectivity for other viruses (Wallinga & Teunis, 2004)), and 2) average presymptomatic infectious period (***pip***) of *4.5* days. This split produces two parameters used in the model:

1) latency rate for the non-infectious period - non-infectious epsilon (***niε***):

$$niε = 1/lnip,$$

and

2) latency rate for the infectious period - latency rate infectious epsilon (***iε***):

$$iε = 1/(lp - lnip).$$

As the Republic of Korea provided high quality, reliable data and conducted a large number of tests during the pandemic, we decided to use Korean proportion of "severe" to diagnosed cases as a base for the probability of developing the severe condition (***pS***), and we set it to *0.01*. We assumed that patients with mild symptoms, in contrast to those in severe condition, are still capable of travelling. For model simplicity, we decided to merge into one compartment all mild and asymptomatic cases.

We decided to set the probability of detection of severe infection (***pDS***) to *0.6*, in order to mimic two obstacles, typically preventing proper accurate diagnosis. Firstly, the majority of patients with a severe course of the disease are either chronically ill or above 60 (Zhou et al., 2020) - their symptoms might be mistaken with those caused by their general health condition, and thus not reported on time. Secondly, the model is supposed to reflect the average illness detection around the globe, which includes many countries with low quality or underfinanced healthcare, where the number of SARS-CoV-2 tests is very limited.

Another parameter of the model, ***pDM*** is the probability of being diagnosed with COVID-19 when having either mild symptoms or an asymptomatic illness course. This parameter depends on previously defined ***pS*** and ***pDS***, as well as the ratio of total diagnosed to undiagnosed cases (***tDR***):

$$pDM = (tDR - pS * pDS) \div (1 - pS).$$

Knowing the limitations of previous modelling attempts (Cowling et al.; Ganyani et al., 2020; Zhang et al., 2020; Chen et al., 2020; Wu, Leung & Leung, 2020; Lin et al., 2020; Kucharski et al., 2020), we decided to test a radically different COVID-19 epidemiologic paradigm, i.e. a significantly lower ***tDR***. It means that in our model, we assume a higher proportion of undetected cases in comparison to other models proposed so far. Taking into account that none of them was capable of providing a plausible global simulation of the pandemic, plus the fact that the potentially low detectability has already been discussed in the literature (Nishiura et al., 2020; Li et al., 2020a; Day, 2020a,b; Kimball et al., 2020), we decided to test such a possibility in simulation by setting the lowest possible ***tDR***. Its relation to ***pDM*** sets its minimum to:

$$tDR > pS * pDS.$$

For previously set ***pS*** and ***pDS*** values, ***tDR*** must be greater than 0.006, thus the value used in our simulation was set to 0.0061. This value means that for 10,000 of COVID-19 cases only 61 are positively diagnosed. The justification for such a strong assumption is based on the following: (i) ***tDR*** reflects the average global detectability of the virus, including countries of low quality of public healthcare; (ii) ***tDR*** reflects the average detectability of the virus during the entire examined period of 154 days that describes early stages of the pandemic; (iii) the percentage of asymptomatic or atypical cases is currently unknown, but small-scale screening tests conducted so far indicate that even 88% of examined diagnosed cases could be asymptomatic (Sutton et al., 2020); and (iv) some of the currently used tests might be faulty e.g. when viral load is small (Pan et al., 2020).

Other important and interconnected parameters required by the model are as follows: the effective contact rate, $\boldsymbol{\beta}$; its reduction level for patients who developed severe symptoms of the disease but were not diagnosed, $\boldsymbol{r\beta}$; and average recovery time since symptoms development $\boldsymbol{\mu}$.
The parameter $\boldsymbol{\beta}$ is derived from the time a host remains infectious, $\boldsymbol{d}$, and the basic reproduction number of the virus, $\boldsymbol{R_0}$:

$$\beta = R_0 \div d \ ,$$

where:

$$d = \mu + pip.$$

The estimation of $\boldsymbol{R_0}$ is a topic widely discussed in the literature, with values ranging from 1.4 to 6.49 ("Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)"; Majumder & Mandl; Zhao et al.; Imai et al., 2020; Read et al., 2020; Liu et al., 2020). However, following the assumption of much higher than the currently suspected ratio of undiagnosed cases, we decided to use in our model a higher rate of transmissibility, yet well within the range of 2-5, modelled for SARS (Wallinga & Teunis, 2004). The $\boldsymbol{R_0}$ value assumed in our model is 4.4.
In our study $\boldsymbol{\mu}$ is derived from a safe quarantine period for diagnosed cases (Woelfel et al., 2020). As the safe quarantine time is estimated to be 10 days (Woelfel et al., 2020), we assumed $\boldsymbol{\mu}$ to last on average for 7 days from symptoms development to recovery. The sum of $\boldsymbol{\mu}$ and previously estimated $\boldsymbol{pip}$ (presymptomatic infectious period) results in $\boldsymbol{d}$ equal to 11.5 days, and $\boldsymbol{\beta}$ equal to 0.38261.
We decided to set $\boldsymbol{r\beta}$ to 0.5, following the assumption for this parameter used in GLEAM modelling of the 2009 influenza outbreak (Balcan et al., 2010). Patients who were diagnosed with COVID-19 are assumed isolated and as such, not spreading the disease any further.

Model compartmentalization

To model the virus spread, we modified the compartmental SIR metapopulation transmission model to represent the nature of the COVID-19 epidemic.
In our model, we used seven different compartments (**Figure 1**).

1. Susceptible population - equal to the general global population. We assume no existing immunity to infection.
2. Latent non-infectious - infected population in the first incubation stage, not yet infectious.
3. Presymptomatic infectious - infected population already infectious, but without developed symptoms.
4. Mild symptoms - joint populations of asymptomatic cases and those with inconspicuous symptoms.
5. Severe symptoms - population infected by SARS-CoV-2 with symptoms affecting their travel ability.
6. Diagnosed - population identified as infected with the SARS-CoV-2 virus. This is the reference line for the model accuracy.
7. Recovered - joint populations of recovered and fatal cases. We assumed recovered individuals cannot be reinfected, although the only evidence so far is for rhesus macaques (Ota, 2020) and WHO is still investigating the issue .

The last step was necessary to avoid the problem of unknown mortality rate of the virus. It should be noted that the currently reported mortality rate only applies to diagnosed cases (CFR, case fatality ratio), and its value still lacks consensus varying from 0.9%-2.1% (Wu et al., 2020).

The true mortality rate (IFR, infected fatality ratio) that takes into account all undiagnosed cases is likely to be much smaller. For China for instance, it has been estimated at 0.66% (Verity et al., 2020). However, even if we assume that we currently detect all COVID-19 cases, it should not exceed 3.4% (WHO estimation) and this already is negligible from the perspective of our model (I.e. artificial increasing of the number of recovered, and thus immune individuals by this amount should not affect the model's final outcome much).

Model running

The prepared model served as an input for 20 runs of GLEAM stochastic simulations. The GLEAM framework uses high-resolution worldwide population data, allowing for the definition of subpopulations according to a Voronoi decomposition of the world surface centered on the locations of International Air Transport Association (IATA)-indexed airports (www.iata.org). Short-range commuting networks for the defined subpopulations are constructed on the basis of data on the commuting patterns of 29 countries in five continents, generalized to a general gravity law for commuting flows, reproducing commuting patterns worldwide.
The stochastic simulation of the pandemic was started on Nov 17th, 2019. Although this date is stated only in non-academic sources, other reports also indicate mid-November as the time of the pandemic outbreak (Li et al., 2020b). The simulation began with a single presymptomatic individual located in Wuhan, China, and the development of the pandemic spread was modelled for 154 days.

Model results processing

A single model run yields two sets of results. The first set is the median value and confidence intervals for the number of individuals per thousand which, at a given day, were moved to each of the compartments (presented in **Figure 1**). The second set is the median value and confidence intervals for a cumulative number of individuals per thousand, entering each of the compartments, until a given day. Both sets of results can be extracted with different resolutions - globally, by hemisphere, continent, country, or the tessellated area surrounding IATA-registered airports.
For areas selected for detailed analyses of model results (i.e. twenty selected countries and the New York area), a cumulative number of transitions into a compartment of interest (i.e. “Diagnosed” and “Mild symptoms” respectively) was multiplied by the area population, and divided by thousand. The display of the model results was limited to the dates for which the experimentally derived data was available.
In order to compare model results with the random screening test from the New York area (Sutton et al., 2020), it was necessary to calculate the average number of undiagnosed mild and asymptomatic cases in this region ($\boldsymbol{c_s}$) for the period covered by the experiment. The median and the lower and upper confidence limits of the number of individuals entering the “Mild symptoms” compartment at any day ($\boldsymbol{n}$) of the simulation is provided by the GLEAM framework ($\boldsymbol{c_{p+m}^n}$, $\boldsymbol{c_{p+l}^n}$, and $\boldsymbol{c_{p+u}^n}$, respectively), while the number of individuals leaving the compartment ($\boldsymbol{c_{p-}^n}$) was estimated similarly as in the framework:

$$c_{p-}^{n} = c_{p}^{n} \div \mu,$$

where c_p^n stands for the average number of infected but not diagnosed asymptomatic or mild symptoms cases at day $\boldsymbol{n}$. For each day ($\boldsymbol{n}$) of the simulation, $\boldsymbol{c_p^n}$ and its lower and upper confidence limits, $\boldsymbol{c_{pl}^n}$ and $\boldsymbol{c_{pu}^n}$, were calculated as:

$$c_p^n = c_p^{n-1} + c_{p+m}^n - c_{p-}^{n-1},$$
$$c_{pl}^n = c_p^{n-1} + c_{p+l}^n - c_{p-}^{n-1},$$
$$c_{pu}^n = c_p^{n-1} + c_{p+u}^n - c_{p-}^{n-1}.$$

The lower and upper confidence limits for $\boldsymbol{c_s}$, respectively $\boldsymbol{c_{sl}}$ and $\boldsymbol{c_{su}}$, used for comparison with confidence interval derived from the screening test, were obtained as:

$$c_{sl} = (\sum_{n=126}^{139} c_{pl}^n)/(139-126), \text{ and}$$
$$c_{su} = (\sum_{n=126}^{139} c_{pu}^n)/(139-126),$$

where $\boldsymbol{n}$ = 126 and $\boldsymbol{n}$ = 139 stand for simulation steps referring to Mar 22nd, 2020, and Apr 2nd, 2020, respectively.

The Excel workbook with performed calculations is provided as Supplementary Workbook 1.

Results

The simulation modelled the pandemic spread for 154 days. The results for all subpopulations around the globe are available in the shared model file (see the **Data Sharing** section below).

As overall data on the pandemic dynamics around the globe is likely to be biased by regions, often considerable in size and population, for which official statistics might be inaccurate, we decided not to compare overall model results with global data. Instead, we limited the analysis of results to twenty countries across five continents which are, in our belief: a) divergent in the proportion of the tested population (as reported in https://ourworldindata.org/coronavirus-testing-source-data), quality of healthcare, and strength of undertaken preventative measures; b) likely to provide the public with real data. We also compared the model outcome for the New York area subpopulation with the results obtained in a random screening test on women admitted for delivery (Sutton et al., 2020).

The obtained 95% confidence intervals of predicted numbers of diagnosed patients were compared with empirical data from twenty countries. In **Figure 2,** we present a percentage difference over time between the number of reported confirmed cases and confidence interval limits for modelled predictions. Positive values state that the model overestimates the number of diagnosed cases; negative values indicate the underestimations of the model; for the observed numbers of diagnosed cases that are within the model's CIs the percentage difference is equal to 0. For most of the selected countries, the model predictions fit well to the observed data, especially in the early stages of the pandemic.

Additionally, **Figure 3** confronts the number of actual confirmed COVID-19 cases with confidence intervals for the modelled number of diagnosed cases in twenty selected countries from five continents for all 154 days of the simulation. Some countries present epidemic dynamics different from the model. However, the direction of these deviations may be explained by the measures overtaken by their governments, their societal response, or the number of tests carried per million of citizens. To show the possible influence of Non-Drug Interventions, we summarized them and marked the dates of their introduction in the country charts. Even though it is difficult to assess the effectiveness of precautions without detailed reports from the country in question, in some cases they seem to explain the observed discrepancies well. For instance the number of detected cases in Japan, Australia and New Zealand is much smaller than predicted by

the model, which might be the result of the ban on flights from China introduced at the beginning of February combined with a geographical isolation of these countries. In contrast, most European countries started introducing preventive actions in March and the potential effects are only slightly visible in the last days of simulation. Similarly, the lack of concordance between model and empirical data in the case of the Republic of Korea may also be caused by the introduced precautions. However, in this case the preventive actions were of a different nature and are not shown in the country chart. Namely, the Republic of Korea introduced what was considered one of the largest and best-organised epidemic control programs in the world, consisting of measures to screen and isolate any infected people, as well as track and quarantine those who contacted them (South Korean Ministry of Health and Welfare).

Furthermore, the empirical data from the last day of the simulation, Apr 19th, 2020, was juxtaposed with the obtained confidence intervals in a single plot for all countries (**Figure 4**). In this plot the number of detected cases is presented as a percentage of the country population, in order to better show the width of the obtained confidence intervals in relation to the entire country population. From **Figure 4** it is visible that model predictions are generally within the same order of magnitude as the observed data. So are the differences between upper and lower confidence limits.

Finally, for the New York area and the time period between March 22nd, 2020 and April 2nd, 2020, we compared the model outcome with the results of a random screening test conducted on women admitted for delivery (Sutton et al., 2020). The cited experimental study revealed that out of 214 women tested, 33 were positively diagnosed with COVID-19, of which 4 had mild symptoms and 29 were asymptomatic. Assuming pregnant women are representative of the entire population and omitting severe COVID-19 cases from calculations (given patients in late pregnancy showing severe symptoms would have been admitted to the hospital earlier), the 95% CI for the true population proportion of mild or asymptomatic cases in the New York area, averaged over the examined period, is 0.11 - 0.21. According to our model, in the same area and averaged over the same period, the 95% CI for this proportion is 0.19 - 0.23. Although the predicted CI is not fully enclosed by the CI of the experimental result, their large overlap suggests high quality and accuracy of the model, started 139 days earlier from a single case on the other side of the globe.

Altogether, our 154-day long simulations of the pandemic seem to reflect the empirical data well. However, as is in the case of any model, this reflection is not perfect. The main reason for the discrepancies between model predictions and the reported number of COVID-19 cases is the fast governmental responses and early introduced precautions, which significantly influence the pace of the disease spread. Such preventive measures, for instance local flight bans, are not included in our simulations. In fact, the model depicts only the “natural” dynamics of the pandemic in the situation when governments do nothing to stop it. This means that in countries where overtaken actions were fast and effective, the model has a tendency to overestimate the number of detected cases.

The second potential reason for the observed discrepancies between model results and empirical data is the increase of the virus detectability in countries where the proportion of tested individuals is larger, leading to higher ***tDR*** than the one assumed in our model. To check this hypothesis, we calculated 95% confidence intervals for the Spearman correlation coefficient between: a) the cumulative number of conducted tests per capita in a country, and b) the percentage difference between the cumulative number of detected cases and the lower or upper confidence limit of the CIs obtained in the model (i.e. if the model underestimates the number of

detected cases, its upper confidence limit is used in calculations, and *vice versa*; if the observed data is within the predictions of the model, the difference is zero). The correlations' CIs were calculated separately for each day of the simulation, if only sufficient data on countries' testing effort were available in the datasource. Missing data on the number of carried tests were interpolated if possible. As some countries started testing earlier than others, the number of datapoints for correlations varies from 4 to 18, depending on the date. The obtained CIs for correlations are plotted on a timeline in **Figure 5**. It is visible that within the used data and with a limited sample size of 4 to 18 countries, it is not possible to state the direction of the correlation and decide if the tested hypothesis is true.

We believe that further modelling efforts, that will include careful parameters' modifications over time in order to better reflect local responses to the pandemic, would greatly improve the accuracy of the simulations, but it is outside of the scope of this work.

Data sharing

The model and the results of the simulation are freely available at https://github.com/freesci/covid19.

The data used to generate **Figure 5** is provided as **Supplementary Table 1**.

The data and calculations used to obtain confidence intervals for the proportion of the mild and asymptomatic cases in the New York area is provided as **Supplementary Workbook 1**.

Discussion

The presented model, due to its stochastic nature, avoids the problem of biased and inaccurate data and provides simulations of the pandemic spread for all the countries around the globe. It also has multiple implications concerning the major characteristics of the COVID-19 pandemic, such as the basic reproduction number of the virus $\boldsymbol{R_0}$ (higher than previously assumed, yet not above the values estimated for other coronaviruses), and the average ratio of diagnosed cases ***tDR*** (much lower than assumed so far, especially for cases expressing mild symptoms and asymptomatic). Such a low ***tDR*** would indicate that the vast majority of the COVID-19 infections are so mild that they pass unnoticed. This is not implausible, considering the fact that there are 1.9 billion children aged below 15 years in the world (27% of the global population) and predominantly (ca. 90%) the course of their infections is mild or asymptomatic (Dong et al., 2020). Additionally, the ***tDR*** used in our model indicates virus detectability averaged over the entire period of 154 days and over all countries in the world. Furthermore, some COVID-19 cases may show atypical symptoms (e.g. diarrhoea) (Gao, Chen & Fang, 2020) which hinder correct diagnosis. Taking all this into account, plus the results of our model and latest reports on the low detectability of the virus (Nishiura et al., 2020; Li et al., 2020a; Day, 2020a,b; Kimball et al., 2020), one may risk a hypothesis that the virus is already more prevalent in the global population than shown in official statistics at the moment, and consequently, its mortality rate is much lower.

To verify this hypothesis, further actions are required. At first, the model should be refined by stochastic fitting of parameters to the observed data. Also the sensitivity analysis of the parameters should be performed. Such a refined model could be used for the analysis of the effectiveness of the applied Non-Drug Interventions and possibly, for the modelling of future NDIs strategies. Secondly, a simulation with the ***tDR*** parameter increasing over time or

diverging geographically might better reflect the actual virus detectability in the course of the pandemic.

Finally, the real spread of the virus should be assessed empirically by conducting a sufficient number of tests on fully random samples as currently, most tests are limited to individuals with strong and typical symptoms. Only after obtaining a solid measurement of the actual prevalence of the virus, can one draw conclusions about its true mortality rate.

We emphasize that our conclusions are a hypothesis based on a single computational model and without empirical verification, but they serve as a platform for further research. At this stage, by no means should they be used as a reason for governmental decisions on lifting the precautions. Even if the true mortality of the virus is indeed lower than announced by the media, to our best knowledge many people remain in the high-risk group (e.g elderly, chronically ill, etc. (Baud et al., 2020)). Lack of population resistance facilitates their contact with the virus and may lead to a rapid increase of severe cases in a short time, as seen for example in Italy (Remuzzi & Remuzzi, 2020), leading to the collapse of the healthcare system, which affects the entire society and results in many additional deaths not related to the virus itself. Careful use and tuning of Non-Drug Interventions, constant balancing of the disease spread and healthcare capacity, protecting the most vulnerable individuals, farsighted anticipation and agility in decision making may altogether be able to minimize the number of deaths without resulting in the global economic breakdown.

Conclusions

Our model implies that the current consensus on the basic reproduction number of SARS-CoV-2 and its prevalence are misestimated. The overall global data on the pandemic dynamics seems strongly biased by large regions where official statistics may not reflect accurately the state of the epidemic, and by the fact that many COVID-19 cases may go unnoticed. The basic reproduction rate of the virus should be confirmed based on reliable data, and its prevalence determined by conducting properly designed screening tests.

References

Balcan D, Gonçalves B, Hu H, Ramasco JJ, Colizza V, Vespignani A. 2010. Modeling the spatial spread of infectious diseases: the GLobal Epidemic and Mobility computational model. *Journal of computational science* 1:132–145.

Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. 2020. Real estimates of mortality following COVID-19 infection. *The Lancet infectious diseases*. DOI: 10.1016/S1473-3099(20)30195-X.

Chen T-M, Rui J, Wang Q-P, Zhao Z-Y, Cui J-A, Yin L. 2020. A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious diseases of poverty* 9:24.

Cowling BJ, Ali ST, Ng TWY, Tsang TK, Li JCM, Fong MW, Liao Q, Kwan MYW, Lee SL, Chiu SS,

Wu JT, Wu P, Leung GM. Impact assessment of non-pharmaceutical interventions against COVID-19 and influenza in Hong Kong: an observational study. DOI: 10.1101/2020.03.12.20034660.

Day M. 2020a. Covid-19: identifying and isolating asymptomatic people helped eliminate virus in Italian village. *BMJ* 368:m1165.

Day M. 2020b. Covid-19: four fifths of cases are asymptomatic, China figures indicate. *BMJ* 369:m1375.

Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*. DOI: 10.1016/S1473-3099(20)30120-1.

Dong Y, Mo X, Hu Y, Qi X, Jiang F, Jiang Z, Tong S. 2020. Epidemiological Characteristics of 2143 Pediatric Patients With 2019 Coronavirus Disease in China. *Pediatrics*. DOI: 10.1542/peds.2020-0702.

Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, Hens N. 2020. Estimating the generation interval for COVID-19 based on symptom onset data. *medRxiv*.

Gao QY, Chen YX, Fang JY. 2020. 2019 Novel coronavirus infection and gastrointestinal tract. *Journal of digestive diseases*. DOI: 10.1111/1751-2980.12851.

Imai N, Cori A, Dorigatti I, Baguelin M, Donnelly CA, Riley S, Ferguson NM. 2020. Report 3: transmissibility of 2019-nCoV. *Reference Source*.

Kimball A, Hatfield KM, Arons M, James A, Taylor J, Spicer K, Bardossy AC, Oakley LP, Tanwar S, Chisty Z, Bell JM, Methner M, Harney J, Jacobs JR, Carlson CM, McLaughlin HP, Stone N, Clark S, Brostrom-Smith C, Page LC, Kay M, Lewis J, Russell D, Hiatt B, Gant J, Duchin JS, Clark TA, Honein MA, Reddy SC, Jernigan JA, Public Health – Seattle & King County, CDC COVID-19 Investigation Team. 2020. Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility - King County, Washington, March 2020. *MMWR. Morbidity and mortality weekly report* 69:377–381.

Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Centre for Mathematical Modelling of Infectious Diseases COVID-19 working group. 2020. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet infectious diseases*. DOI:

10.1016/S1473-3099(20)30144-4.

Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich NG, Lessler J. 2020. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of internal medicine*. DOI: 10.7326/M20-0504.

Lavezzo E, Franchin E, Ciavarella C, Cuomo-Dannenburg G, Barzon L, Del Vecchio C, Rossi L, Manganelli R, Loregian A, Navarin N, Abate D, Sciro M, Merigliano S, Decanale E, Vanuzzo MC, Saluzzo F, Onelia F, Pacenti M, Parisi S, Carretta G, Donato D, Flor L, Cocchio S, Masi G, Sperduti A, Cattarino L, Salvador R, Gaythorpe KAM, Imperial College London COVID-19 Response Team, Brazzale AR, Toppo S, Trevisan M, Baldo V, Donnelly CA, Ferguson NM, Dorigatti I, Crisanti A. 2020. Suppression of COVID-19 outbreak in the municipality of Vo, Italy. *medRxiv*:2020.04.17.20053157.

Lin Q, Zhao S, Gao D, Lou Y, Yang S, Musa SS, Wang MH, Cai Y, Wang W, Yang L, He D. 2020. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases* 93:211–216.

Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J. 2020a. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*. DOI: 10.1126/science.abb3221.

Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine* 27. DOI: 10.1093/jtm/taaa021.

Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, Chaillon A. 2020b. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *Journal of medical virology*. DOI: 10.1002/jmv.25731.

Majumder M, Mandl KD. Early Transmissibility Assessment of a Novel Coronavirus in Wuhan, China. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3524675.

513 Nishiura H, Kobayashi T, Suzuki A, Jung S-M, Hayashi K, Kinoshita R, Yang Y, Yuan B,
514 Akhmetzhanov AR, Linton NM, Miyama T. 2020. Estimation of the asymptomatic ratio of novel
515 coronavirus infections (COVID-19). *International journal of infectious diseases: IJID: official*
516 *publication of the International Society for Infectious Diseases*. DOI: 10.1016/j.ijid.2020.03.020.
517 Ota M. 2020. Will we see protection or reinfection in COVID-19? *Nature reviews. Immunology*. DOI:
518 10.1038/s41577-020-0316-3.
519 Pan Y, Long L, Zhang D, Yan T, Cui S, Yang P, Wang Q, Ren S. 2020. Potential false-negative nucleic
520 acid testing results for Severe Acute Respiratory Syndrome Coronavirus 2 from thermal inactivation
521 of samples with low viral loads. *Clinical chemistry*. DOI: 10.1093/clinchem/hvaa091.
522 Read JM, Bridgen JRE, Cummings DAT, Ho A, Jewell CP. 2020. Novel coronavirus 2019-nCoV: early
523 estimation of epidemiological parameters and epidemic predictions. *medRxiv*.
524 Remuzzi A, Remuzzi G. 2020. COVID-19 and Italy: what next? *The Lancet* 395:1225–1228.
525 South Korean Ministry of Health and Welfare.Coronavirus disease 19(COVID-19). *Available at*
526 *http://ncov.mohw.go.kr/en/* (accessed April 25, 2020).
527 Statement on the meeting of the International Health Regulations (2005) Emergency Committee regarding
528 the outbreak of novel coronavirus 2019 (n-CoV) on 23 January 2020. *Available at*
529 *https://www.who.int/news-room/detail/23-01-2020-statement-on-the-meeting-of-the-international-*
530 *health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-*
531 *(2019-ncov)* (accessed April 23, 2020).
532 Statement on the second meeting of the International Health Regulations (2005) Emergency Committee
533 regarding the outbreak of novel coronavirus (2019-nCoV). *Available at https://www.who.int/news-*
534 *room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-*
535 *(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov)* (accessed
536 March 20, 2020).
537 Steven Sanche, Yen Ting Lin, Chonggang Xu, Ethan Romero-Severson, Nick Hengartner, Ruian Ke.
538 2020. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus

2. *Emerging Infectious Disease journal* 26. DOI: 10.3201/eid2607.200282.

Streeck H, Hartmann G, Exner M, Schmid M. 2020. Preliminary result and conclusions of the COVID-19 case cluster study (Gangelt Municipality). *Available at: https://www.land.nrw/sites/default/files/asset/document/zwischenergebnis_covid19_case_study_gangelt_en.pdf* (accessed April 29, 2020).

Sutton D, Fuchs K, D'Alton M, Goffman D. 2020. Universal Screening for SARS-CoV-2 in Women Admitted for Delivery. *The New England journal of medicine*. DOI: 10.1056/NEJMc2009316.

Tizzoni M, Bajardi P, Poletto C, Ramasco JJ, Balcan D, Gonçalves B, Perra N, Colizza V, Vespignani A. 2012. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC medicine* 10:165.

Tong Z-D, Tang A, Li K-F, Li P, Wang H-L, Yi J-P, Zhang Y-L, Yan J-B. 2020. Potential Presymptomatic Transmission of SARS-CoV-2, Zhejiang Province, China, 2020. *Emerging infectious diseases* 26. DOI: 10.3201/eid2605.200198.

Van den Broeck W, Gioannini C, Gonçalves B, Quaggiotto M, Colizza V, Vespignani A. 2011. The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC infectious diseases* 11:37.

Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, Cuomo-Dannenburg G, Thompson H, Walker P, Fu H, Dighe A, Griffin J, Cori A, Baguelin M, Bhatia S, Boonyasiri A, Cucunuba ZM, Fitzjohn R, Gaythorpe KAM, Green W, Hamlet A, Hinsley W, Laydon D, Nedjati-Gilani G, Riley S, van-Elsand S, Volz E, Wang H, Wang Y, Xi X, Donnelly C, Ghani A, Ferguson N. 2020. Estimates of the severity of COVID-19 disease. *medRxiv*:2020.03.09.20033357.

Wallinga J, Teunis P. 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American journal of epidemiology* 160:509–516.

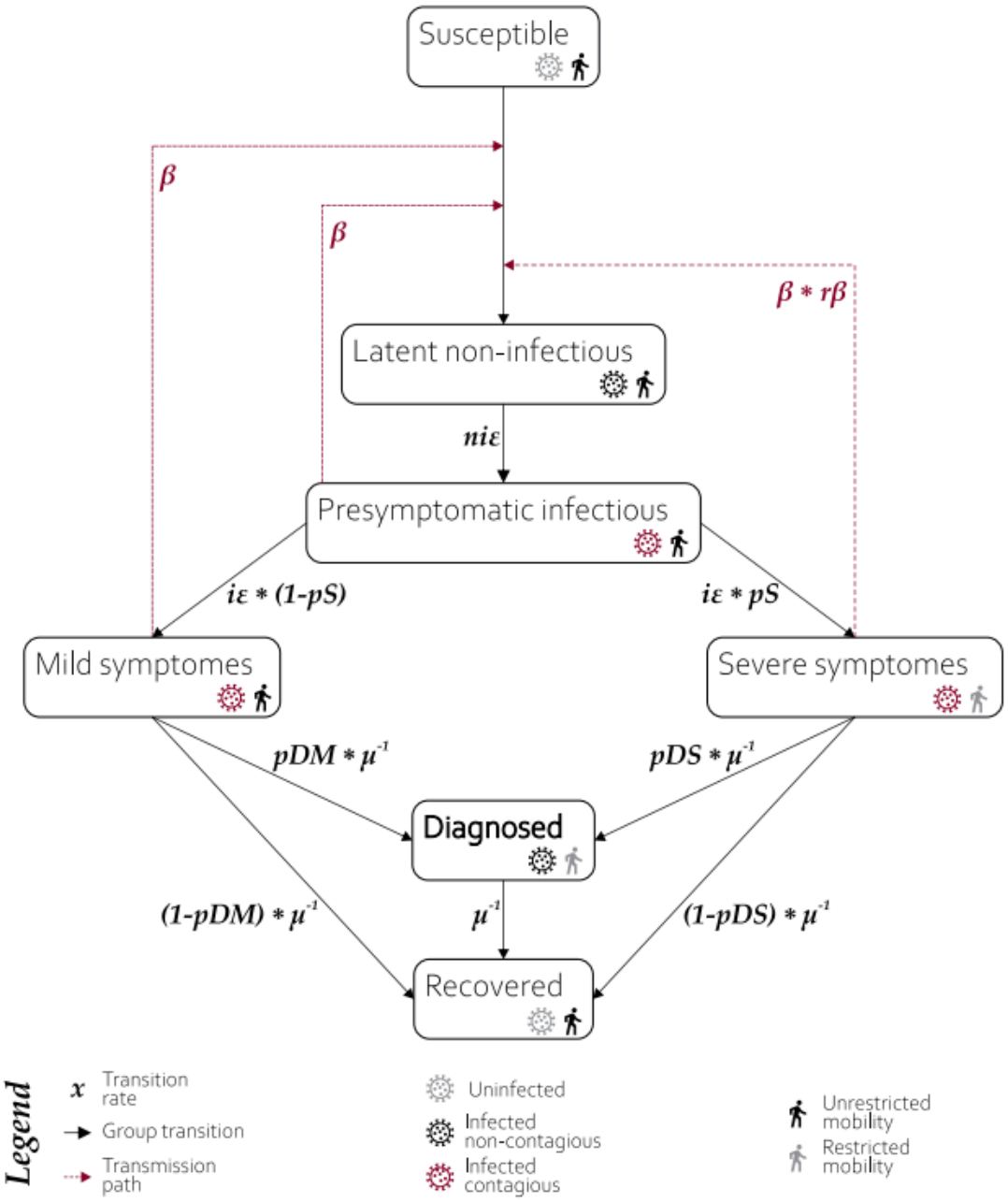
Woelfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Mueller MA, Niemeyer D, Vollmar P, Rothe C, Hoelscher M, Others. 2020. Clinical presentation and virological assessment of hospitalized cases of coronavirus disease 2019 in a travel-associated transmission cluster. *medRxiv*.

565 Wu JT, Leung K, Bushman M, Kishore N, Niehus R, de Salazar PM, Cowling BJ, Lipsitch M, Leung
566 GM. 2020. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan,
567 China. *Nature medicine* 26:506–510.
568 Wu JT, Leung K, Leung GM. 2020. Nowcasting and forecasting the potential domestic and international
569 spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*
570 395:689–697.
571 Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D. 2020. Estimation of the reproductive number of novel
572 coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A
573 data-driven analysis. *International journal of infectious diseases: IJID: official publication of the*
574 *International Society for Infectious Diseases* 93:201–204.
575 Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, Lou Y, Gao D, Yang L, He D, Wang MH.
576 Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in
577 China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. DOI:
578 10.1101/2020.01.23.916395.
579 Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, Guan L, Wei Y, Li H, Wu X,
580 Xu J, Tu S, Zhang Y, Chen H, Cao B. 2020. Clinical course and risk factors for mortality of adult
581 inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*. DOI:
582 10.1016/S0140-6736(20)30566-3.

583

584

585

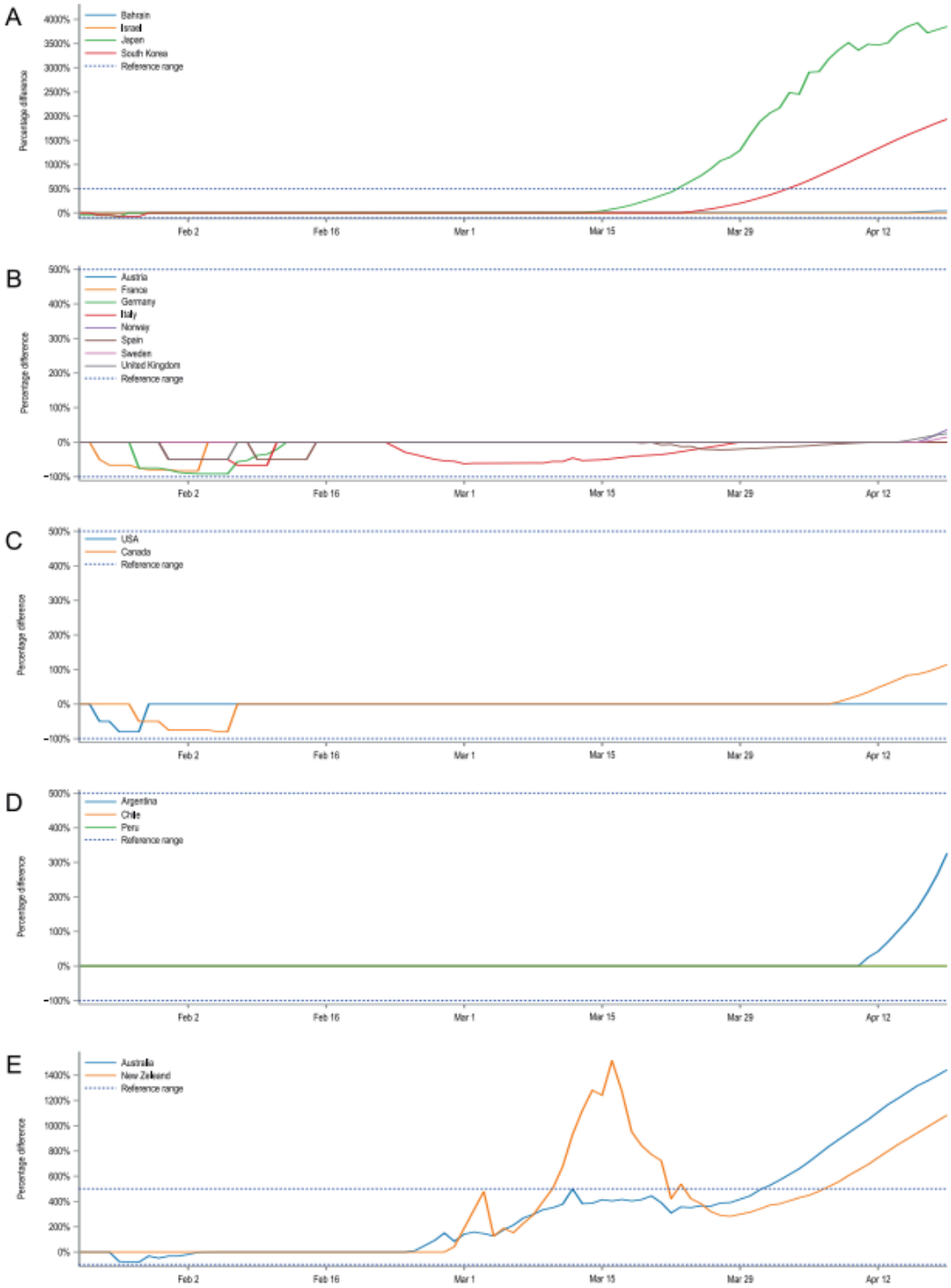


586

587 *Figure 1*: Structure of compartments used in modelling.
588 A susceptible individual in contact with a person: 1) presymptomatic, 2) who developed mild
589 symptoms, or 3) who developed severe symptoms, may contract the infection at rate ***β***, ***β*** or ***rβ*β***,
590 respectively, and enter the latent non-infectious compartment where he is infected but not yet
591 infectious. During the non-infectious period, each individual has a probability of ***niε*** of becoming

592 presymptomatic infectious. The presymptomatic cases have probability $i\varepsilon$ of developing severe
593 or mild COVID-19 symptoms, with probabilities pS and $1-pS$, respectively. The transition from
594 both symptomatic groups occurs at μ rate. Individuals who developed severe symptoms do not
595 travel within and between modelled subpopulations and may be either diagnosed with probability
596 pDS, or recover with probability of $1-pDS$. Individuals whose mild (or non-existent) symptoms
597 are not stopping them from traveling may be diagnosed with probability pDM or recover with
598 probability $1-pDM$. The diagnosed individuals are considered isolated and effectively non-
599 contagious and recover with rate μ. The recovery does not discriminate between true recovery
600 and fatal cases.
601

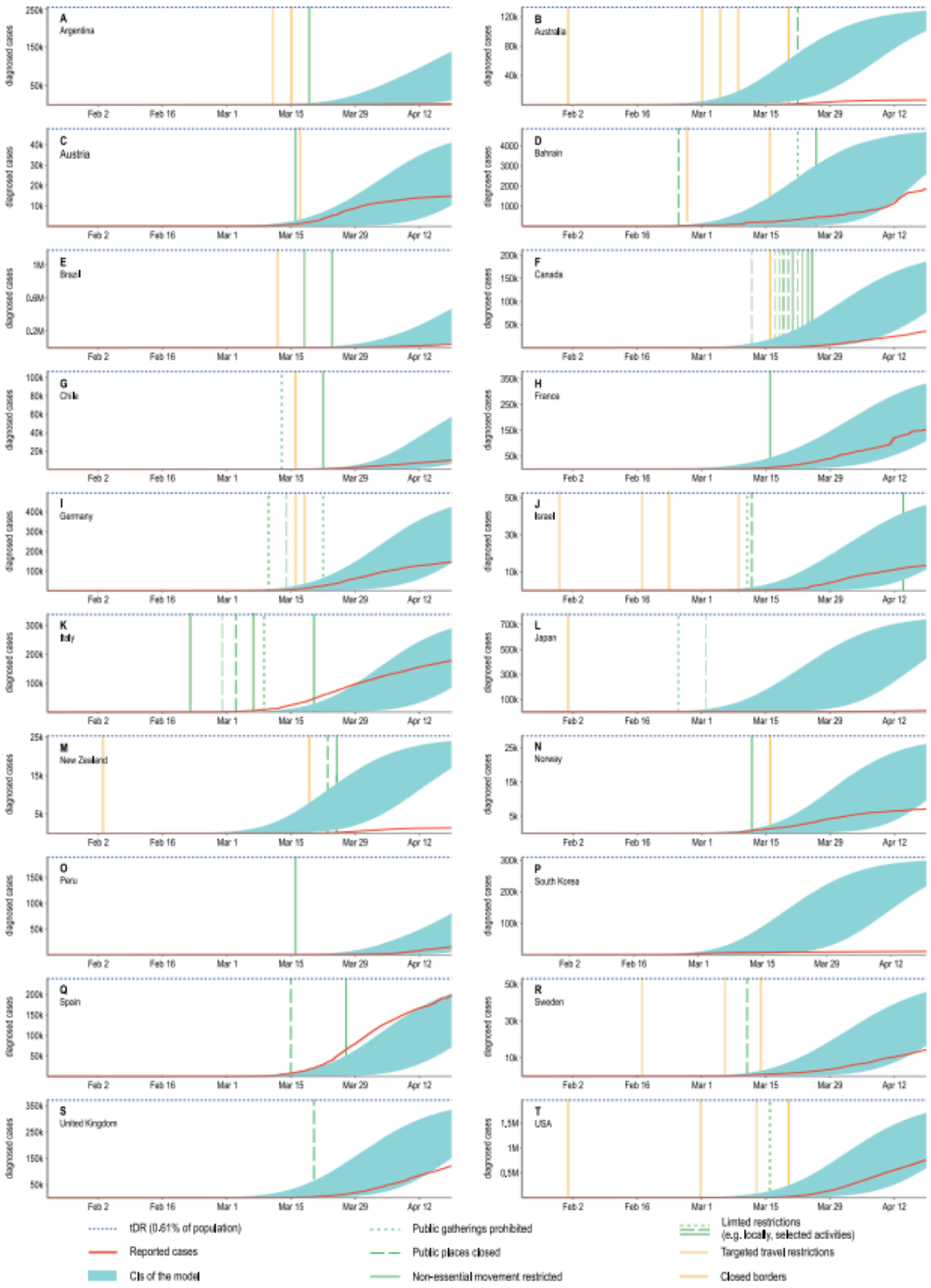
602



603

Figure 2: Percentage difference over time between the number of reported cases and confidence intervals' limits for modelled predictions.
Positive values state that the model overestimates the number of diagnosed cases, negative values indicate the underestimations of the model. Observed numbers of cases that are within the model CIs are equal to 0. For clarity, country plots were grouped by continents and presented in five plots: **A**: Asia, **B**: Europe, **C:** North America, **D:** South America, **E:** Australia and Oceania. The large discrepancies for Japan, Australia, New Zealand and the Republic of Korea are putatively caused by the fast and pronounced reaction of their governments and early introduced NDIs which are not reflected in our model.

615
616



617

Figure 3: An overlay of modelled confidence intervals for the predicted cumulative number of diagnosed cases and the actual reported values shown for twenty selected countries. The y-axes show the absolute number of diagnosed cases and due to different country populations are not unified. To facilitate comparisons a blue, dotted line was added as a reference indicating the 0.61% of the total population of a country. This value is the same as the ***tDR*** parameter used in our model reflecting the assumed ratio of detected to undetected cases. The confidence intervals obtained in our model will approach this value asymptotically. For most countries, observations agree well with model predictions. The observed discrepancies are most likely due to introduced NDIs which are not included in our model. The precautions were categorized and the dates of their introductions were marked on the plots with vertical lines, however, the assessment of their effectiveness is beyond the scope of this research.

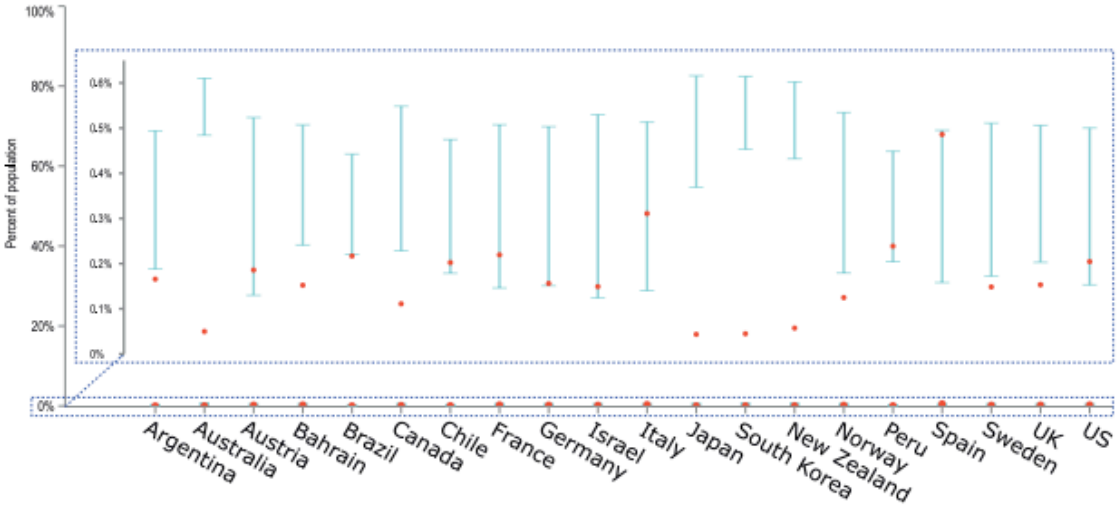
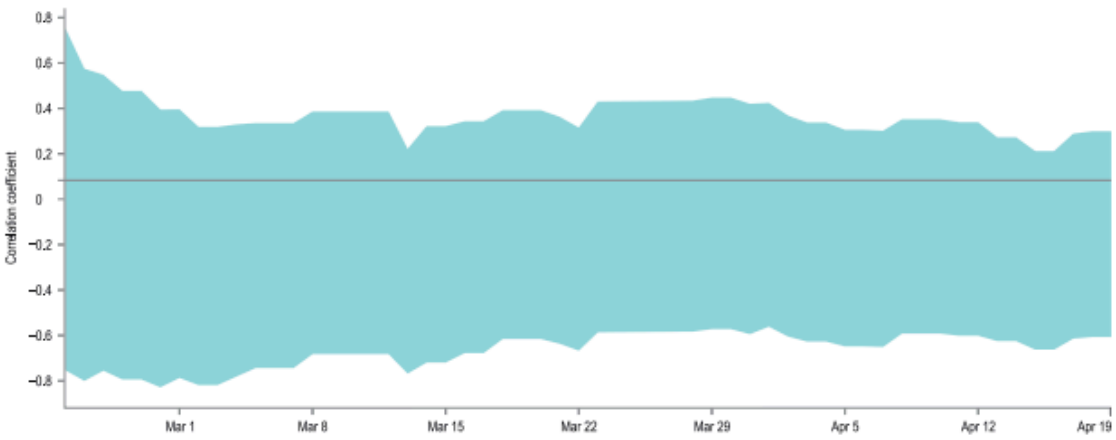


Figure 4: An overlay of modelled confidence intervals with the empirical data on the number of diagnosed cases for the last day of the simulation, Apr 19th, 2020.
The cumulative number of diagnosed cases is presented as a percentage of the country population, facilitating comparisons between countries. Model predictions are generally within the same order of magnitude as the observed data and obtained CIs are relatively narrow despite a limited number of iterations (20).

642



643
644

645 *Figure 5*: Correlation between presented model accuracy and per-country testing effort.
646 An overlay of 95% confidence intervals for the Spearman correlation coefficients between: 1)
647 the interpolated, cumulative number of conducted tests per capita in a country, and 2) the
648 percentage difference between the actual number of detected cases and the lower or upper
649 confidence limit of the CIs obtained in the model. The correlations were calculated separately for
650 each day of the simulation if only sufficient data on testing effort is available. As some countries
651 started testing earlier than others, the number of data points for correlations varies from 4 to 18,
652 depending on the date. The width of the obtained CIs and the values of their limits indicate that
653 with the avaialble data it is not possible to state the direction of the analyzed correlations.
654

655

656

Parameter	Value	Description	Source/Derivation
R_0	*4.4*	Reproduction number for SARS-CoV-2	Literature-based: assumed on the basis of infectivity rates of other coronaviruses.
μ	*7 days*	Average recovery time since symptoms development	Literature-based
β	*0.38261*	Transmission rate	$\mu \div R_0$
rβ	*0.5*	Reduction in transmission rate resulting from the undiagnosed development of severe COVID-19 symptoms	Literature-based
lp	*5.6 days*	Average latency period	Literature-based
lnip	*1.1 days*	Average latent non-infectious period	Literature-based: assumed on basis of non-infectious period of Influenza A/H1N1
pip	*4.5 days*	Average presymptomatic infectious period	$lp - lnip$
niε	*0.9(09)*	Probability of transition from ***lnip*** to ***pip*** state	$1 \div lnip$
iε	*0.2(2)*	Probability of transition from presymptomatic to symptomatic state	$1 \div (lp - lnip)$
pS	*0.01*	Probability of developing severe COVID-19 symptoms	Literature-data: The most often reported ratio of severe to mild symptoms
pDS	*0.6*	Probability of being diagnosed when expressing severe COVID-19 symptoms	Assumed, taking into account that: 1. In elderly patients, COVID may be easily misdiagnosed, 2. Most of the countries in the world do not have sufficiently efficient healthcare systems.
tDR	*0.0061*	Rate of diagnosed SARS-CoV-2 infected individuals	Value representing the smallest possible detectability $> pS * pDS$
pDM	*0.00(01)*	Probability of being diagnosed when presenting mild or none COVID-19 symptoms	$(tDR - pS * pDS) \div (1 - pS)$

657
658

659 Table 1. List of parameters used for the models and their respective values.