CIRRELT

CIRRELT-2020-34

Modeling the COVID-19 Pandemic: A Sensitivity Analysis on Input Data Using Agent-Based Transportation Simulation

Ouassim Manout
Idriss El-Megzari
Francesco Ciari

September 2020

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

Modeling the COVID-19 Pandemic: A Sensitivity Analysis on Input Data Using Agent-Based Transportation Simulation

Ouassim Manout[1,2,*], Idriss El-Megzari[2], Francesco Ciari[1,2]

[1] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)

[2] Department of civil, geological and mining engineering, Polytechnique Montreal. idriss.el-megzari; francesco.ciari@polymtl.ca

Abstract. The COVID-19 pandemic has rampaged through the world causing hundreds of thousands of fatalities, millions of infections, and acute economic and social consequences. In this context, epidemiological models have been put forward as key tools to help public health authorities and policy makers cope with the pandemic. In some cases, these models have proved to be useful aid-decision tools. Nonetheless, in several instances, these models still face several shortcomings and challenges. First, most popular epidemic models still rely on an aggregate representation of the population. Second, the reliability and accuracy of these models in the presence of biased input data is yet to be tested. In this research, we address these two challenges by conducting a sensitivity analysis on the impact of bias in input data, using a transportation agent-based model (MATSIM), and a MATSIM-dependent epidemic model, EPISIM. Our findings stress that epidemic predictions are sensitive towards bias in input data. The underestimation of confirmed infections and the overestimation of the rate of fatalities in official public health reports, bias modeling outcomes. We found that when such biased data are used to train EPISIM, predictions of severe cases and infections are often overestimated. This overestimation bias conveys a false picture of the dynamics of the pandemic. The policy implications of these findings are of interest to both researchers and policy makers.

Keywords: COVID-19, epidemic, data bias, modeling, agent-based, transportation, MATSIM, EPISIM.

Acknowledgements. This work is carried out thanks to the open-source model EPISIM and the TU Berlin team that developed it. The authors acknowledge their valuable work. The authors also acknowledge the indirect contribution of Dr. Pieter Fourie to this paper. At an early stage of this research, the authors have discussed the research ideas of this paper with Dr. Fourie.

---

* Corresponding author: ouassim.manout@polymtl.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2020

Introduction

In few months, the current COVID-19 pandemic has rampaged through the world causing nearly 17 million confirmed infections and a global death toll of more than 675,000 as of August first. Such rapid spread seems to be the product of the inherent characteristics of the virus and of the globalized and interconnected world. The Spanish flu pandemic, in 1918-19 is probably the only other example of a pandemic at such scale in the recent history of humanity. The conditions under which that happened were remarkably different from now, as the world was much less connected and populations were weakened by years of war. At that time, epidemic models did not exist. The Kermack and McKendrick (1927) model, largely considered as the progenitor of current epidemic models, came only around one decade later. Now, almost a century later, epidemiology still largely relies on relatively simple mathematical models.

As the COVID-19 pandemic progressed, the modeling efforts to predict its diffusion multiplied, some of them trying to address the limitations of the existing approaches. One of the main shortcomings of these models is the lack, or the simplistic representation of the spatial dimension of the infection and the diffusion vector. Transportation models have faced similar challenges when modeling mobility. These models consider space to represent how individuals move from one location to another using the transportation network. Historically, models were based on aggregated representations of travel dynamics, like modeling trip flows from zone to zone. Nowadays, completely disaggregated models where individuals are explicitly represented, emerged as a viable alternative. A special sub-class of such models is the agent-based transportation simulation.

The transportation agent-based approach is particularly interesting in the context of epidemic modeling. This approach has proved its contribution to finely modeling individual characteristics and detailed activity plans. This contribution can benefit epidemic models. First, the whereabouts of each agent are known as well as the type of activities. Second, a statistically representative picture of each individual and of the total population of the simulated area is provided. The first contribution implies that we can track individuals, their contacts in space and time and for each activity type. This provides the lacking spatial dimension of infection models. Also, each agent is described by individual characteristics that can be relevant to the spread of the virus.

In light of the above, embedding an epidemic model into an agent-based simulation would potentially overcome several limitations; but also would introduce, at least, a new one. Whereas standard epidemic models can, a priori, be fed with few observations like the number of infections, an agent-based epidemic model will require more complete and detailed data: socio-demographics of infected agents, their spatial distribution, etc. This paper is not the first proposing to use agent-based simulation to model pandemic (Frias-Martinez et al., 2011, Müller (2020); Perez and Dragicevic, 2009). The paper builds on the work of Müller (2020) to understand the impact of the imperfect knowledge of the infection cases on pandemic predictions.

The next section reviews epidemic models and the potential bias of imperfect input data. The research questions are stated in section 3. Data and methods are detailed in section 4. Section 5 presents the results. Sections 6 and 7 discuss and conclude this research by showing the policy implications of its findings.

Background

On the importance of epidemic models

From the first days of the COVID-19 outbreak, epidemiological research has played a critical role. These studies have pushed forward our understanding of the ongoing pandemic and have helped decision makers monitor and control for its dire consequences. In this context, prediction models are key tools. These models predict the evolution of the pandemic and its main future dynamics: future number of infections, number of deaths and critical cases, the peak period, or the coming pressure on the health system. On the basis of these predictions, public health authorities have recommended unprecedented policy interventions, like physical distancing or total lockdowns.

Consequently, much of COVID-19 research has been put in the study and the design of sophisticated epidemic prediction models (Chinazzi et al., 2020; Currie et al., 2020; Giordano et al., 2020; Lin et al., 2020; Liu et al., 2020; Meehan et al., 2020). Nevertheless, the accuracy and reliability of these models have sometimes been overlooked (Holmdahl and Buckee, 2020; Jordan, 2020). In some instances, epidemic models have delivered conflicting conclusions and raised questions and critics about their reliability.[1]

One of the most recurring causes of modeling flaws of epidemic models is bias in input data (Fauci et al., 2020; Holmdahl and Buckee, 2020; Jordan, 2020; Lu et al., 2020; Roda et al., 2020). Most COVID-19 models rely, to a certain degree, on observations to model infection dynamics and extrapolate their future evolution. Data accuracy and reliability is therefore a crucial issue in the calibration and validation of these models in order to operate as decision aid tools. Most models often rely on official counts and reports. In several instances, these data sources have been shown to be biased and not accurate (Krantz and Rao, 2020; Reis et al., 2020).

Bias in epidemic modeling

As demonstrated by the COVID-19 pandemic, public health authorities have limited knowledge of past and ongoing infection cases, especially during the first weeks of the outbreak. Often, official counts systematically underreport mild-symptom infections and focus on more severe cases (Basu, 2020; Fauci et al., 2020; Jordan, 2020; Lau et al., 2020). On one hand, not all COVID-19 cases are detectable. According to a recent meta-analysis review on asymptomatic cases, the authors found an asymptomatic incidence between 6% and 41%, with an average proportion of 16% (Byambasuren et al., 2020). The magnitude of this variance depicts how official counts of infections might be biased. Also, in the case of COVID-19 outbreak, first infection cases have often gone unnoticed for weeks if not months. In Europe, early infected regions have officially reported their first cases in late February 2020. Current studies ascertain that, in these regions, first cases were in fact introduced earlier, in December 2019 (Deslandes et al., 2020). On the other hand, in different parts of the world, testing capacities are still insufficient to detect all cases and are often targeting specific disease

[1] See for example this media coverage on COVID-19 prediction models (Begley, 2020; Bui et al., 2020; Piper, 2020).

epicenters or vulnerable communities, making these tests non-random and their results biased. Furthermore, to the previous errors in data gathering, adds up the bias of overreporting the rates of severe cases and fatalities. Patients with serious complications are often hospitalized and their cases reported. Counts on these cases are consequently more accurate than mild-symptom or asymptomatic cases. This induces a non-random bias in pandemic data. In this regard, different authors have shown that naive pandemic indicators, like the fatality ratio, are often biased (Fauci et al., 2020).

Despite these biases in official data, public health authorities and academics often resort to using these data sources to model infection spread. In this regard, we stress the need to question the accuracy, the reliability and the sensitivity of these models towards data biases. We especially target two sources of bias:

- Non-random missing data in official counts.
- Quantity of past data used to train the model.

The non-random missing data bias relates to the relatively unequal data quality and accuracy of official infection counts. The non-random bias in data reporting has already been demonstrated to exaggerate the severity and the fatality rate of the COVID-19 disease (Basu, 2020; Fauci et al., 2020; Jordan, 2020). If one can argue that, in these circumstances, it is better to exaggerate than to downplay the unknown severity of the infection, one can also argue that such bias might endorse inefficient public health measures and policy decisions, like denying hospital access to non-COVID-19 emergencies to avoid shortages in hospital beds (Vergano et al., 2020).

To model the spread of infection, one needs data. Often epidemic data are only available weeks after the outbreak. This delay in data availability has two contrasting effects: (i) a need for short delays to make model predictions available to policy makers at the onset of the pandemic; (ii) a need for longer delays to have more calibration data and better reliability of predictions. Ideally, one would prefer to use a model with no delay from the first day of the outbreak. Practically, models need observation data for training and validation, and these data come only at a price of a delay in observation. Inevitably, there should be a trade-off in choosing a delay. The greater the delay, the more data the model can use to replicate the past, the better the predictions are. At the same time, the lesser the delay, the more rapidly predictions can enlighten policy makers.

Epidemic models and the agent-based approach

Infection models can roughly fall in two categories: statistical and epidemiological models. Statistical models are mathematical forecasting tools used to predict infection trends given past observed data. These models are often trained on time-series data, like infection cases and fatality records. Statistical models have no epidemiological background and are data-driven. Consequently, these models are simple to implement, need few parameters to be set, but they offer limited modeling capabilities. Epidemiological models, on the other hand, implement an epidemiological explanatory model to explain the spread of infection and the evolution of the disease. The most popular epidemiological approach in infection modeling is the Susceptible, Infectious, Recovered (SIR) approach (Kermack and McKendrick, 1927). This approach models different infection stages or compartments and transitions from one stage to another. Relatively to statistical models,

epidemiological models offer interesting modeling capabilities, but also require serious calibration efforts (Currie et al., 2020; Holmdahl and Buckee, 2020). From a theoretical point of view, epidemiological models are also more accurate than statistical ones.

One interesting epidemiological modeling approach is based on the agent-based paradigm. Agent-based models (ABM) have long been used in research, and recently introduced in epidemiology (Ajelli et al., 2010; Frias-Martinez et al., 2011; Kai et al., 2020; Müller, 2020; Perez and Dragicevic, 2009). This approach allows combining the explanatory power of epidemiological models and the fine granularity of agent-based tools. In this regard, they offer the opportunity to model agent-to-agent virus transmission by accounting for behavioural aspects of agents, their social networks, their mobility and activity patterns, and more importantly, by considering their heterogeneity: age, gender, spatial location, etc.

In transportation, agent-based models have long been studied and recently popularized by numerous applications. In this regard, the open-source framework MATSIM, is one of the most used and maintained ABM in transportation research. The Multi-Agent Transport Simulation (MATSIM) framework is a state-of-the-art agent-based transportation model (Horni et al., 2016). It has been developed for more than 20 years, and successfully applied to different urban contexts and at different spatial scales. At the core of MATSIM, there is a synthetic population. The synthetic population is a detailed description of main characteristics of each individual of the population of interest. These characteristics include various socio-demographic data like age, sex, or household members and detailed daily activity plans of each agent. In MATSIM, agents concurrently try to optimize their daily plans that eventually brings the system to an equilibrium that tends to replicate observed data.

MATSIM outcome is valuable information to epidemiological modeling, since it tracks each individual and models its interactions, in space and time, with others. In this context, a new MATSIM-dependent tool, dubbed EPISIM, has recently been introduced to model infection spread (Müller, 2020). In this paper, we make use of MATSIM and EPISIM and their flexibility to study the sensitivity of pandemic predictions towards bias in input data.

Research questions

Bias in input data is likely to induce biased modeling predictions and endorse policy decisions with, at best, unpredictable consequences. In this regard, it is vital to assess the reliability and accuracy of epidemic models. In this paper, we tackle two research questions:

1. Are COVID-19 agent-based models sensitive towards bias in input data?
2. What is the impact of such sensitivity on infection predictions?

To answer these questions, we rely on a transportation agent-based model, MATSIM, coupled with an epidemic model, EPISIM.

Methods

We study the sensitivity of EPISIM, a MATSIM-dependent agent-based epidemic model, towards bias in input data, and especially towards underreported official counts and delays in modeling infection spread.

EPISIM

Epidemic Simulation, or EPISIM, is an open-source multi-day agent-based pandemic modeling framework (Müller, 2020). EPISIM implements a SIR-like epidemiological modeling approach (Kermack and McKendrick, 1927). But, in contrast with the conventional SIR approach that is based on differential equations to move individuals from one state to another, EPISIM uses a probabilistic model. It has been recently developed to model the ongoing COVID-19 outbreak in Berlin, Germany (Müller, 2020). To model the spread of the virus, EPISIM uses MATSIM events. MATSIM events record the basic steps each agent takes to perform its daily activities. These data describe when and where agents take part of common activities, like work, leisure or travel, with whom and for how long. Such information is valuable to infection modeling.

Based on MATSIM events, EPISIM computes for each agent the chances of getting infected by others using a probabilistic model. In the SIR approach, agents are either susceptible, i.e. exposed to the virus, infected or recovered, i.e. no longer susceptible. All agents, except some, are assumed to be susceptible in the first days of simulation. On each day of simulation, susceptible agents bear the risk of getting infected when interacting with other agents in public and private places. After some time, infected agents recover from the infection and are assumed immune for the rest of the simulation. To trigger the infection process, a predefined number of randomly chosen infectors, or patients-zero, are introduced in the population during the first days.

Infection model

Each time an agent takes part of an activity that involves others, the agent might infect or get infected. The infection process is probabilistic, and the probability of infection is determined by a parametric model. This model accounts for different variables relevant to the viral transmission (eq. 1).

$$\begin{cases} P_{A->B} = 1 - \exp(\theta \times R) \\ R = \Delta_{A-B} \times T_{A-B} \times V_A \times S_B \times (1 - M_A) \times (1 - M_B) \geq 0 \\ \theta < 0 \end{cases} \tag{1}$$

$P_{A->B}$ is the probability of agent A to infect agent B.

R is the risk agent B runs when interacting with agent A. The higher the risk, the higher is the probability of getting infected. $R \in [0, +\infty[$.

Δ_{A-B} is the physical distance between agents A and B. $\Delta \in]0, +\infty[$. Since MATSIM does not model the physical proximity of agent, we assume that $\Delta_{A-B} = 1$ for all interactions.

T_{A-B} is the interaction duration. $T \in [0, +\infty[$. This time is provided by MATSIM events.

V_A is the viral load, i.e. how contagious is A. The viral load depends on the disease status of the agent. Serious cases have a viral load of 1. Other contagious cases are assumed to have a viral load of 0.7. $V \in [0, 1]$.

S_B is the susceptibility of getting infected. The susceptibility translates the uneven propensity of agents to become infected, when exposed to the same viral load. The susceptibility depends on the health condition of agents. We use the age as a proxy for this. In this paper, we assume elderly people ($age \geq 65$) as being more susceptible ($S = 1$) than younger agents. $S = 0.5$ for agents with an age between 10 and 64, and $S = 0.1$ for agents with less than 10 years of age. $S \in [0, 1]$.

$M_{A/B}$ is a risk-reduction factor to account for individual protective measures, like wearing a mask. $M \in [0, 1]$. In this research paper, we assume that no protection measure is used ($M = 0$).

θ is a viral transmission calibration parameter to fit EPISIM outcomes to observed data: infections and severe cases. θ is computed using a multi-objective minimization of a mean squared logarithmic error (MSLE) function.

By design, MATSIM is not endowed with contact tracing capabilities. The precise information on interactions between specific agents is not known. Modeling these interactions is challenging since they depend on various unobserved factors. In this regard, EPISIM assumes that each agent undertaking an activity in a given facility, like work-office, shopping center or transit vehicle, is assumed to interact with a predefined number of randomly selected agents present in the same facility at the same time.

Transition model

EPISIM implements a transition model that extends the SIR framework with seven different transition states (figure 1). The transition from one state to another is either time-dependent or probability-dependent. Time-dependent transitions ensure a deterministic evolution of the disease at the beginning and at the end of the infection. Probability-dependent transitions introduce variability in disease severity among infected agents. At each probabilistic transition (from contagious to showing symptoms, for example), there is a chance that the agent remains in its current disease stage (contagious for example) till recovery. Time and probability-dependent transition values have been set using observed COVID data (Müller, 2020). All infected agents recover after 14 days of infection. Deliberately, the transition model does not include a fatality compartment after the critical stage.

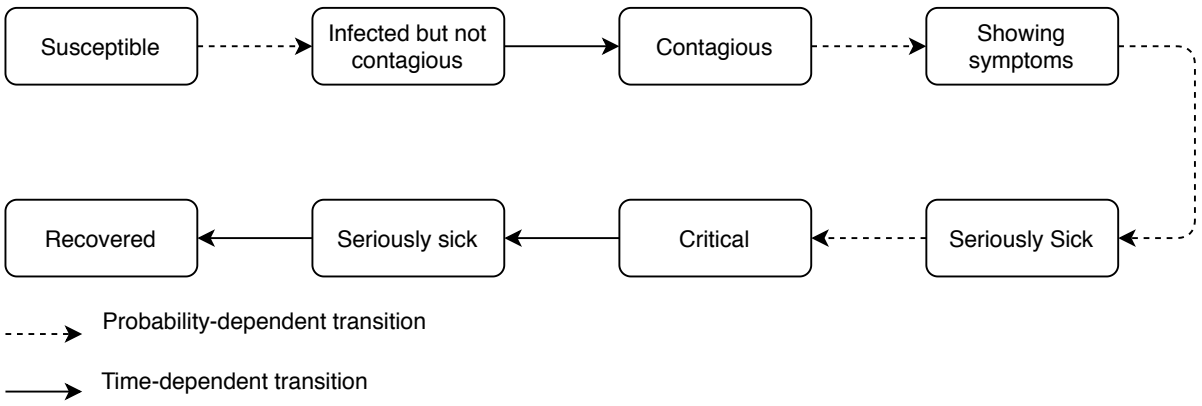

Figure 1: EPISIM transition states

In contrast with EPISIM assumptions, we assume that seriously sick and critical cases are infectious with the highest viral load possible. If these cases are often hospitalized in facilities where safety rules should be scrupulously followed, data show that, in some cases, hospitals have become active infection epicenters given the concentration of infections.

Sensitivity study

While it is possible to study the sensitivity of infection predictions with numerous epidemic models, the suggested approach seems to be one of the most convenient. Thanks to the explicit representation of individuals and of infection transmission and transition, this allows making reasonable assumptions on which cases go unnoticed or underreported, and how such individuals will contribute to the spread of the virus, given their characteristics, and how this influences the dynamics of the pandemic as a whole.

To study the sensitivity of pandemic predictions to bias in input data, we start by setting up an ideal reference pandemic situation. In this reference situation, all infection cases are reported, including asymptomatic ones. The deterministic evolution of the COVID-19 pandemic is also known for each day in the future. Hereafter, this ideal situation is called the reference model and its outcome is considered as observation. Based on this hypothetical case study with ideal knowledge, we introduce different biased versions of reference data to account for errors in observations. In particular, two types of errors are considered:

- Non-random missing data in official counts.
- Delays to get official data for model training.

To account for the non-random missing bias in official counts, we deliberately deteriorate the quality of inputs and train new pandemic models using these erroneous data. The bias we introduce in observations mimics the varying degree of accuracy of pandemic data. Especially, we assume that 90% of seriously sick cases and 95% of critical cases are known and reported. We also assume that only a limited proportion of non-serious or mild-symptom infections is reported to health authorities. For the sake of simulation, we consider this proportion as being a random variable following a uniform distribution from 10% to 90% of known infection

cases. Furthermore, 20% of cases are considered asymptomatic, of which only 2% are detectable. All the rest of cases are falsely considered as being susceptible, i.e. not yet infected.

To account for the delay in observations caused by data gathering, we fit different epidemic models with a delay ranging from 2 to 12 weeks.

Research protocol

Two parameters are used to define the biased pandemic models (figure 2):

- The rate of observation (R_{obs}): proportion of confirmed cases in total infections. It ranges from 10% to 90% of observed data, i.e. reference data.
- The delay in observations ($Delay$): number of recording days of infections. It ranges from 15 to 90 days after the effective start of the pandemic.

Practically, for each variant of the model, the reference model is replayed and its simulation is halted at a specific breaking point day to account for the delay in observations (figure 2). A missing-data factor is then introduced in the reference model to account for the imperfect knowledge of pandemic cases (R_{obs} in figure 2). At the breaking point, an online new calibration is conducted to update the infection model, θ (eq. 1) and transition probabilities (figure 1), to replicate the new biased observations. Afterwards, the simulation is resumed and new pandemic predictions are made using the biased model (dashed curves in figure 2).

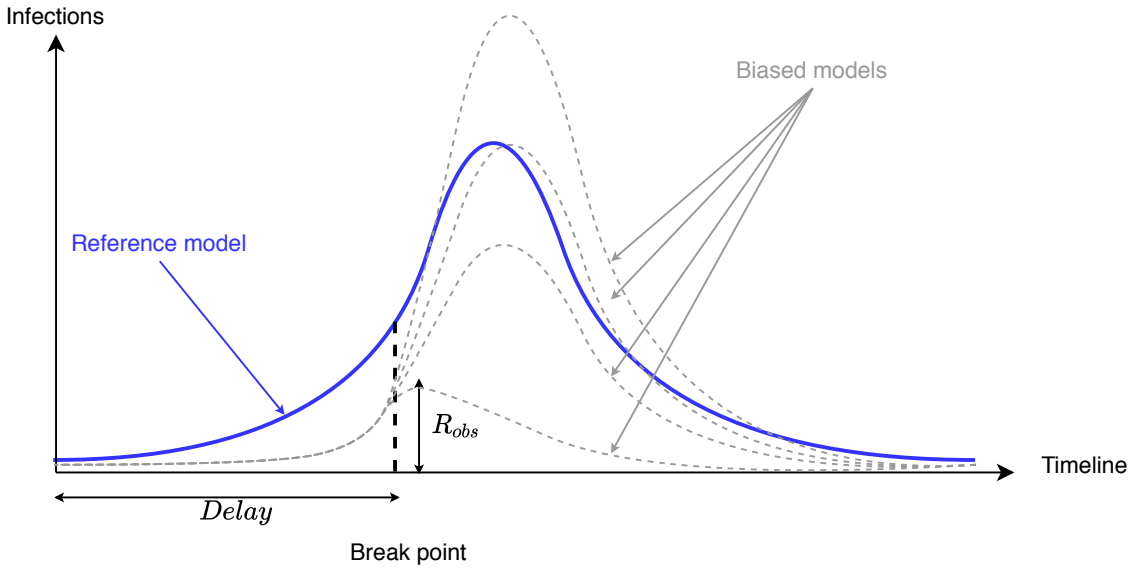

Figure 2: Research protocol for the definition of biased models of pandemic spread

To limit the computation burden, we choose $R_{obs} \in [0.1, 0.9]$ with a step size of 0.1, and $Delay \in [15, 90]$ with a step size of 15 days. 54 ($9R_{obs} \times 6Delays$) different prediction models are defined and compared to

the reference one.

Data

The sensitivity of pandemic predictions is demonstrated in the case of Sioux Falls in South Dakota, USA. The Sioux Falls scenario has been used as a benchmark in several transportation studies. It has been enriched and adapted to MATSIM (Chakirov and Fourie, 2014). This scenario describes one-day activity plans of 84,110 agents. Two aggregate activity types are considered: work and secondary activities. Activities are located at the building level. To perform their trip chains, agents can use cars, public transit or walk as travel modes.

An enriched version of this scenario is used for pandemic modeling. Three hospitals with a total capacity of 900 beds are created with an assumption of 1 bed per 100 capita (an upper limit according to the Organisation of Economic Co-operation and Development (2020)). 100 Nursing home facilities with a capacity of 50 beds are also added. For each facility a capacity-proportional number of patients (70% of bed capacity) and healthcare workers (20% of bed capacity) are assigned. We assume that 70% of the total bed capacity is always occupied by patients and elderly residents. 5,630 patients and nursing home residents are added to the original Sioux Falls population and assigned health facilities as a home location. Healthcare workers are assigned to these facilities using a spatial gravity model. Besides patients and healthcare workers, these facilities also attract visitors. All the residents of health facilities are assigned a maximum susceptibility factor (eq. 1). These improvements of the Sioux Falls synthetic population are meant to better model the infection spread and spatial clusters of the disease.

The pandemic Sioux Falls scenario is a complete hypothetical case study. Input data and modeling outcomes should not, in any case, be compared with real or taken as real data.

Results

EPISIM outcomes describe infection dynamics in detail: who infected whom, when and where. Also, EPISIM tracks the evolution of the disease stage of each infected agent. In this research, we use various pandemic indicators to illustrate our findings: infection curves, maximum number of active infections, maximum number of active seriously sick and critical cases, peak day of infections, of seriously sick and of critical cases.

The infection curve is a key pandemic modeling outcome. It describes the timely evolution of infections, the general trend and peak period. The maximum number of active cases is also an informative epidemiological outcome. When computed for seriously sick and critical cases, this indicator is often compared to available hospital beds to detect potential shortages in health provision and pressure on the health system. Based on these outputs, we compare observed data, i.e. reference outcomes, to the new models we have defined.

Rate of observation

Underreported and non-random missing information in input data have a significant impact on pandemic predictions. The extent of this impact depends on the proportion of missing information (R_{obs}) (figure 3, figure 4). As expected, the more data are missing, the higher the bias in predictions. This is true for all pandemic indicators; but especially significant for predictions of critical and seriously sick cases.

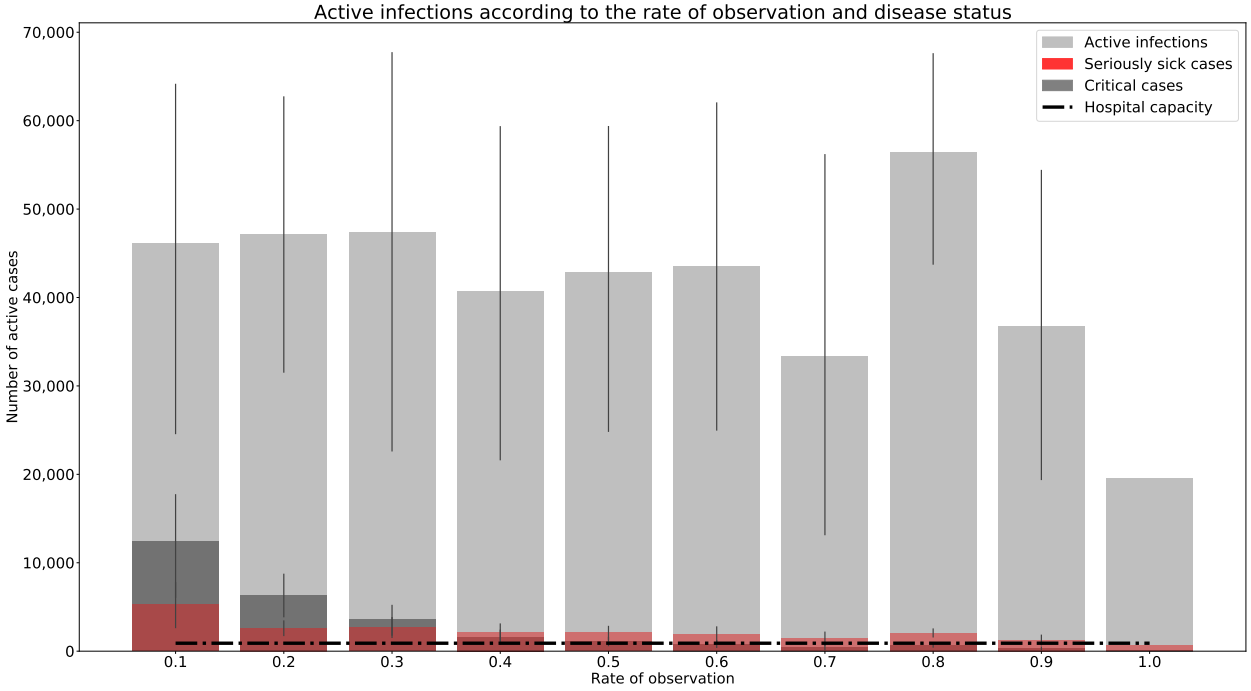

Figure 3: Distribution of total active infections according to the rate of observation and disease status. For each rate of observation, 6 different values corresponding to different delays in observation are depicted by a confidence interval at the top of the corresponding bar. Reference data have a rate of observation of 1.

Pandemic models trained on missing data overestimate the infection spread and virulence. The extent of the overestimation bias is significant for all infection types. When only 10% of infection cases are confirmed, i.e. reported, the maximum cumulative active infections are overestimated, on average, by nearly 3-fold. The average overestimation reaches its highest levels with predictions of severe cases (figure 4). Maximum active seriously sick cases are overestimated by 9-fold. Maximum active critical cases are overestimated by more than 100-fold. The extent of this bias remains high even when a relatively low rate of missing data is used. When 90% of total mild-symptom cases are reported and used in pandemic models, active infections are still overestimated by 2-fold factor, on average. Active seriously sick and critical cases are, respectively, overestimated by 2 and 3-fold factors.

For each rate of observation, confidence intervals in figure 3 show how important the variance of the outcomes depending on the delay in observation. Also, for severe cases, findings suggest that the bias in pandemic outcomes follows a decreasing exponential trend. The marginal impact of the non-random missing data is

more important for rates of observation under 50% than for higher rates.

The impact of the rate of observation on the infection peak period is less important than with predictions of the number of active infections. On average, when the rate of observations increases from 10% to 90%, we found a maximum 8-day shift of the peak relatively to the reference data. When the rate of observation increases, the infection peak shifts to the right.

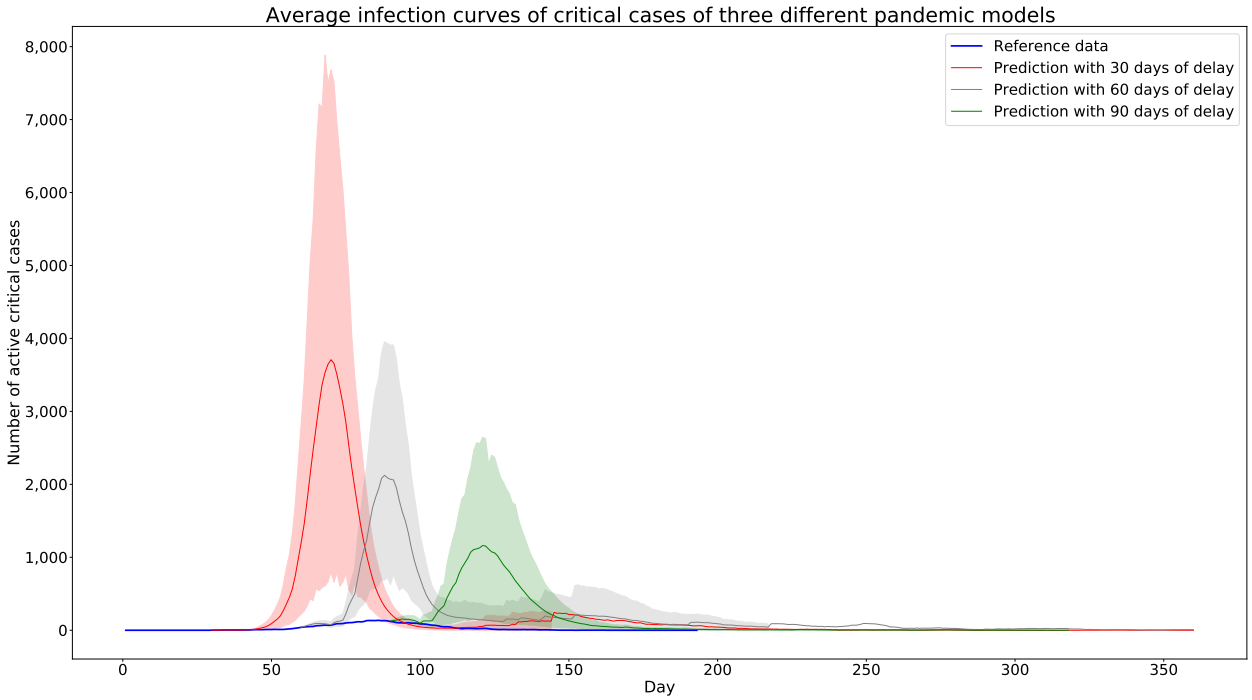

Figure 4: Average infection curves critical cases of three epidemic models with a delay of 30, 60 and 90 days. The confidence interval around the curves relates to different rates of observation.

Delay in observations

Delays in observation have a substantial effect on epidemic predictions (figure 5). A delay of 15 days induces an overestimation of 3-fold of the maximum active infections. This overestimation error decreases when the delay increases. For delays greater than 75 days, predicted maximum active infections are close to observations. With a delay of 90 days, pandemic models underestimate, on average, the maximum number of active infected cases by -9.5%. Nonetheless, for severe cases, i.e. seriously sick and critical cases, delays in observation still overestimate their incidence whatever the delay.

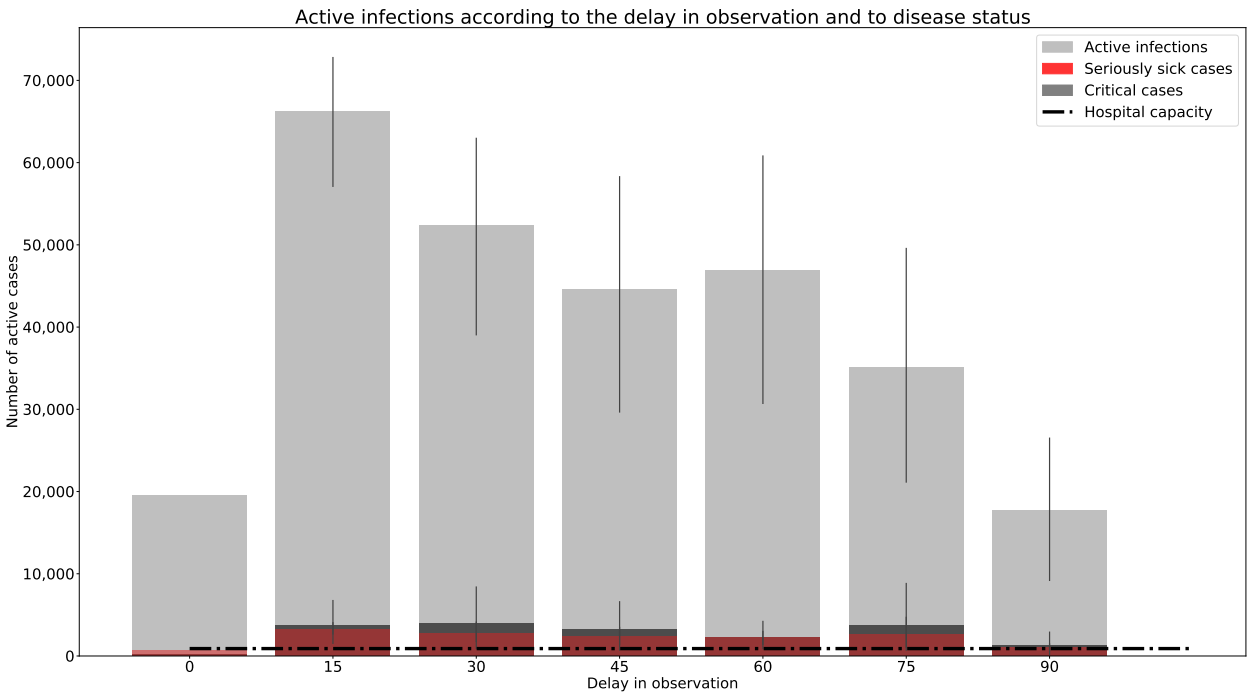

Figure 5: Distribution of total active infections according to the delay in observation and disease status. For each delay, 9 different values corresponding to different rates of observation are depicted by a confidence interval at the top of the corresponding bar. Reference data have a delay of 0.

Delays also bias the prediction of the peak period of infections. When the delay is less than 45 days, pandemic models predict a relatively much earlier infection peak than reference data. Results show that the peak period of infections shifts to the right when the delay increases from 15 to 90 days.

Discussion

This research highlights one main finding: epidemic predictions are significantly sensitive towards errors in input data. This research also confirms the findings of preceding papers on the contribution of agent-based transportation models to pandemic modeling.

This paper confirms the findings of Müller (2020), on the contribution of MATSIM-EPISIM to the modeling of the COVID-19 infection. This approach offers the opportunity to model, at a fine scale, the infection spread from agent-to-agent by using daily travel plans and their details in space and time. This also allows to analyse the dynamics of the infection at a fine spatial scale to identify potential pandemic clusters and to adopt appropriate policy interventions.

The transportation agent-based approach extends epidemic modeling capabilities and overcomes some of its shortcomings. By considering detailed activity plans of agents and their interactions in space and time, the modeling of infection spread replicates more accurately reality. Agents need not to be similar as assumed

by the standard SIR model. Agents have individual characteristics that vary among the population and that may influence the spread of the virus (like age or spatial location). Also, each agent has a specific contact network that is embedded in space and time, instead of a fixed rate of contact for all individuals as assumed by the SIR assumptions of a well-mixed population and constant rate of contact. In this regard, we think that the agent-based approach is a promising step towards more realistic and, hopefully, more accurate epidemic models and predictions. A more complete and rich demonstration of this is given in Müller (2020).

As of the accuracy of epidemic models, our findings suggest that EPISIM predictions are sensitive towards bias in input data and delays in observation. The extent of this bias is significant and questions the reliability of these outcomes, in particular, and the predictions of epidemic models and the SIR approach, in general, especially when endorsing vital public health decisions.

The impact of bias in input data is threefold:

1. Overestimation of the probabilities of transition to severe disease stages.
2. Bias in the calibration parameter of the infection model, i.e. θ in eq. 1.
3. Overestimation of the number of susceptible individuals during the spread of the pandemic.

Results demonstrate that biased and underreported input data are detrimental to the accuracy of epidemic predictions. Predictions of the maximum number of active infections or their cumulative number are often overestimated; but to a lesser degree when compared with severe cases. Predictions of severe cases are key information to public health authorities and hospitals to manage the pressure on the health system. These predictions are found to be overestimated many folds when the rate of confirmed cases is underreported relatively to severe cases. This overestimation is due to the rates and probabilities of transition that are at the core of the SIR approach. Transition probabilities from one disease stage to another (figure 1) are computed using observed data and are normalized using the number of confirmed infections. These probabilities are inevitably overestimated when confirmed cases are underreported. Such overestimation has also been reported in different COVID-19 studies that rely on the SIR approach and on ratios to move infected people from one disease compartment to the next (Grant, 2020; Mallapaty, 2020; Roda et al., 2020).

With non-random missing data, the calibration parameter of the infection model is biased. In our case, this induces an overestimation of the virulence of the Coronavirus. The value of θ has been found to be often greater than its true value in the reference model. Moreover, the underestimation of infections means also an overestimation of susceptible individuals, namely more opportunities for the virus to spread. This contributes to the overestimation of total infections in the population and the shift of the peak period.

Results show that the delay in observations has a substantial impact on predictions of total infections and peak periods, and a relatively, little impact on predictions of severe cases. The delay variable controls the size of data used to train the epidemic model. The shorter the delay, the less data can be used for calibration. Hence, short delays induce non-robust statistical modeling parameters and uncertain predictions. Furthermore, with the overestimation bias induced by the non-random missing data, infection parameters are biased and overestimated. Therefore, with shorter delays, EPISIM produces non-robust and overestimated infection curves with earlier peak period than observations. With longer delays, more data become available for

calibration and less bias is introduced in predictions.

Conclusion

Policy implications

Results confirm the sensitivity of EPISIM predictions, and more generally, of the SIR approach towards bias in input data. The policy implications of these findings are important, especially in the current situation where public health authorities and policy makers are in need of scientific advice. Our results show that models, despite their sophistication, might be subject to bias. One common bias that has plagued most current COVID-19 models, is the errors in reporting and counting infection cases. Official counts of confirmed infections of COVID-19, often, underreport the magnitude of the pandemic. We show that this bias is detrimental to infection predictions and to the reliability of their conclusions. In several instances, these errors induce a systematic overestimation of infections and severe cases. Should these predictions be used to endorse health measures, they will likely support inappropriate policies with unpredictable consequences on society.

Findings also highlight the critical need for sensitivity analysis of epidemic models to ensure their reliability and accuracy. In the first months of the current pandemic, little attention has been put in this task, and impactful policy decisions have been, sometimes, adopted, partly, on the basis of these models. In some instances, some popular models have been invalidated by reality which raised critics and concerns on the overall validity of models (Bui et al., 2020; Piper, 2020). The best answer to these reasonable concerns is to put forward the limits of our models and, at least, to conduct sensitivity and uncertainty analysis on their outcomes.

This paper, among others (Frias-Martinez et al., 2011; Müller, 2020; Perez and Dragicevic, 2009), calls for the development of a research agenda on epidemiological agent-based modeling. This approach is promising and can boost epidemiological research to cope with some of the challenges we have failed to meet during the COVID-19 pandemic. Various research questions are still to be addressed. The calibration and validation of these tools are among these questions. In this regard, advances made in transportation research and the ABM approach can benefit epidemiology, as EPISIM and MATSIM demonstrate this.

Finally, this paper stresses the critical importance of epidemic data quality and reliability. Bias in official counts, like non-random omission or underreports, is detrimental to modeling efforts and policy decisions. If it is not possible to ensure bias-free data, it would be, at least, advisable to acknowledge that official counts might be underreported and to include corresponding uncertainty margins. This also calls for the development of more robust models towards underestimation errors. One way to achieve this is by conducting Monte-Carlo-like simulations to account for errors in input data. Moreover, online-calibration can address the delay trade-off between the need for long time-series data to train the model and the need for rapidly operational models. Online calibration strategies update the model as new data become available and, subsequently, update corresponding confidence intervals.

Limitations

One major limitation of this work comes from the online calibration used in the research protocol. At each breaking point, we conduct an online calibration to update the infection model to fit biased observations. This calibration process raises two concerns:

- Computation time: the search for the minimum of the multi-objective function is costly. In our paper, we bounded the search space for the θ parameter and also limited the size of the candidates to 20 values. We can reasonably overcome these limits with high performance computing. Due to resource and time constraints, this was not possible in the current paper.
- Objective function: we use a mean squared logarithmic error function for the calibration. This function penalizes underestimations more than overestimations. This property might be of interest to epidemic models that may favor overestimation of infection incidence over underestimation. Nonetheless, this has partly contributed to the overestimation bias in our findings. This limitation can be removed by testing another symmetrical calibration function like the standard mean square error.

Another limitation to this work is the use of a one-day activity plan for EPISIM, instead of multi-day plans. In our experiments, we assume that the Sioux Falls population performs the same daily plans each day of the week, including weekends. This is far from reality. Nevertheless, in our research the emphasis is more on the sensitivity of EPISIM outcomes than on modeling the exact infection dynamics. The use of multi-day plans would likely have contributed to modify the infection spread speed; but our findings will likely hold true.

Finally, this research can be enhanced by a study of the sensitivity of epidemic predictions when policy interventions and mask wearing are adopted. In this paper, we assume that no intervention policy is enforced in order to untangle the impacts of bias in data from other effects.

Contribution of authors

The authors confirm contribution to the paper as follows: study conception and design: Prof. Ciari, Ph.D Manout; data collection: Ph.D Manout; analysis and interpretation of results: Ph.D Manout, Prof. Ciari, MSc. El-Megzari; draft manuscript preparation: Ph.D Manout, Prof.Ciari, Msc. El-Megzari. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgements

This work is carried out thanks to the open-source model EPISIM and the TU Berlin team that developed it. The authors acknowledge their valuable work. The authors also acknowledge the indirect contribution of Dr. Pieter Fourie to this paper. At an early stage of this research, the authors have discussed the research ideas of this paper with Dr. Fourie.

References

1. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london Series A, Containing papers of a mathematical and physical character. 1927;115(772):700–21.

2. Perez L, Dragicevic S. An agent-based approach for modeling dynamics of contagious disease spread. International Journal of Health Geographics [Internet]. 2009;8(1):50. Available from: https://doi.org/10.1186/1476-072x-8-50

3. Frias-Martinez E, Williamson G, Frias-Martinez V. An agent-based model of epidemic spread using human mobility and social network information. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing [Internet]. IEEE; 2011. Available from: https://doi.org/10.1109/passat/socialcom.2011.142

4. Müller MAN Sebastian Alexander AND Balmer. Mobility traces and spreading of covid-19 [Internet]. Technische Universität Berlin; 2020. Available from: http://dx.doi.org/10.14279/depositonce-9835

5. Liu Z, Magal P, Seydi O, Webb G. A COVID-19 epidemic model with latency period. Infectious Disease Modelling [Internet]. 2020;5:323–37. Available from: https://doi.org/10.1016/j.idm.2020.03.003

6. Lin Q, Zhao S, Gao D, Lou Y, Yang S, Musa SS, et al. 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 [Internet]. 2020 Apr;93:211–6. Available from: https://doi.org/10.1016/j.ijid.2020.02.058

7. Giordano G, Blanchini F, Bruno R, Colaneri P, Filippo AD, Matteo AD, et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in italy. Nature Medicine [Internet]. 2020 Apr;26(6):855–60. Available from: https://doi.org/10.1038/s41591-020-0883-7

8. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science [Internet]. 2020 Mar;eaba9757. Available from: https://doi.org/10.1126/science.aba9757

9. Meehan MT, Rojas DP, Adekunle AI, Adegboye OA, Caldwell JM, Turek E, et al. Modelling insights into the COVID-19 pandemic. Paediatric Respiratory Reviews [Internet]. 2020 Jun; Available from: https://doi.org/10.1016/j.prrv.2020.06.014

10. Currie CS, Fowler JW, Kotiadis K, Monks T, Onggo BS, Robertson DA, et al. How simulation modelling can help reduce the impact of COVID-19. Journal of Simulation [Internet]. 2020 Apr;14(2):83–97. Available from: https://doi.org/10.1080/17477778.2020.1751570

11. Holmdahl I, Buckee C. Wrong but useful what covid-19 epidemiologic models can and cannot tell us. New England Journal of Medicine [Internet]. 2020 Jul;383(4):303–5. Available from: https://doi.org/10.

1056/nejmp2016822

12. Jordan MI. On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. Harvard Data Science Review [Internet]. 2020 Jun; Available from: https://doi.org/10.1162/99608f92.f01ee285

13. Piper K. This coronavirus model keeps being wrong. Why are we still listening to it? [Internet]. Vox. 2020 [cited 2020 Jul 28]. Available from: https://www.vox.com/future-perfect/2020/5/2/21241261/coronavirus-modeling-us-deaths-ihme-pandemic

14. Bui Q, Katz J, Parlapiano A, Sanger-Katz M. What 5 Coronavirus Models Say the Next Month Will Look Like [Internet]. The New York Times. 2020 [cited 2020 Jul 28]. Available from: https://www.nytimes.com/interactive/2020/04/22/upshot/coronavirus-models.html

15. Begley S. Influential Covid-19 model shouldn't guide U.S. policies, critics say [Internet]. STAT. 2020 [cited 2020 Jul 28]. Available from: https://www.statnews.com/2020/04/17/influential-covid-19-model-uses-flawed-methods-shouldnt-guide-policies-critics-say/

16. Lu FS, Nguyen AT, Link NB, Lipsitch M, Santillana M. Estimating the early outbreak cumulative incidence of COVID-19 in the United States: Three complementary approaches. 2020 Apr; Available from: https://doi.org/10.1101/2020.04.18.20070821

17. Fauci AS, Lane HC, Redfield RR. Covid-19 navigating the uncharted. New England Journal of Medicine [Internet]. 2020 Mar;382(13):1268–9. Available from: https://doi.org/10.1056/nejme2002387

18. Roda WC, Varughese MB, Han D, Li MY. Why is it difficult to accurately predict the COVID-19 epidemic? Infectious Disease Modelling [Internet]. 2020;5:271–81. Available from: https://doi.org/10.1016/j.idm.2020.03.001

19. Krantz SG, Rao ASS. Level of underreporting including underdiagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. Infection Control & Hospital Epidemiology [Internet]. 2020 Apr;41(7):857–9. Available from: https://doi.org/10.1017/ice.2020.116

20. Reis RF, Melo Quintela B de, Oliveira Campos J de, Gomes JM, Rocha BM, Lobosco M, et al. Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in south korea, italy, and brazil. Chaos, Solitons & Fractals [Internet]. 2020 Jul;136:109888. Available from: https://doi.org/10.1016/j.chaos.2020.109888

21. Lau H, Khosrawipour V, Kocbach P, Mikolajczyk A, Ichii H, Schubert J, et al. Internationally lost COVID-19 cases. Journal of Microbiology, Immunology and Infection [Internet]. 2020 Jun;53(3):454–8. Available from: https://doi.org/10.1016/j.jmii.2020.03.013

22. Basu A. Estimating the infection fatality rate among symptomatic COVID-19 cases in the United States. Health Affairs [Internet]. 2020 Jul;39(7):1229–36. Available from: https://doi.org/10.1377/hlthaff.2020.

00455

23. Byambasuren O, Cardona M, Bell K, Clark J, McLaws M-L, Glasziou P. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. 2020 May; Available from: https://doi.org/10.1101/2020.05.10.20097543

24. Deslandes A, Berti V, Tandjaoui-Lambotte Y, Alloui C, Carbonnelle E, Zahar J, et al. SARS-CoV-2 was already spreading in france in late december 2019. International Journal of Antimicrobial Agents [Internet]. 2020 Jun;55(6):106006. Available from: https://doi.org/10.1016/j.ijantimicag.2020.106006

25. Vergano M, Bertolini G, Giannini A, Gristina GR, Livigni S, Mistraletti G, et al. Clinical ethics recommendations for the allocation of intensive care treatments in exceptional, resource-limited circumstances: The italian perspective during the COVID-19 epidemic. Critical Care [Internet]. 2020 Apr;24(1). Available from: https://doi.org/10.1186/s13054-020-02891-w

26. Ajelli M, Gonçalves B, Balcan D, Colizza V, Hu H, Ramasco JJ, et al. Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. BMC Infectious Diseases [Internet]. 2010 Jun;10(1). Available from: https://doi.org/10.1186/1471-2334-10-190

27. Kai D, Goldstein G-P, Morgunov A, Nangalia V, Rotkirch A. Universal Masking is Urgent in the COVID-19 Pandemic: SEIR and Agent Based Models, Empirical Validation, Policy Recommendations. arXiv:200413553 [physics, q-bio] [Internet]. 2020 Apr [cited 2020 Jul 29]; Available from: http://arxiv.org/abs/2004.13553

28. Horni A, Nagel K, Axhausen K, editors. Multi-agent transport simulation matsim. London: Ubiquity Press; 2016.

29. Chakirov A, Fourie PJ. Enriched Sioux Falls scenario with dynamic and disaggregate demand. 2014; Available from: http://hdl.handle.net/20.500.11850/80996

30. Organisation of Economic Co-operation, Development. Health Care Resources [Internet]. 2020 [cited 2020 Jul 24]. Available from: https://stats.oecd.org/index.aspx?DataSetCode=HEALTH_REAC

31. Mallapaty S. How deadly is the coronavirus? Scientists are close to an answer. Nature [Internet]. 2020 Jun;582(7813):467–8. Available from: https://doi.org/10.1038/d41586-020-01738-2

32. Grant A. Dynamics of COVID-19 epidemics: SEIR models underestimate peak infection rates and overestimate epidemic duration. 2020 Apr; Available from: https://doi.org/10.1101/2020.04.02.20050674