Quantifying the over-estimation of predicted pandemic curves caused by population clustering

Mathias L. Heltberg,[1,2,3,4,*] Christian Michelsen,[1,2] Emil S. Martiny,[2] Lasse Engbo Christensen,[5] Mogens H. Jensen,[2] Tariq Halasa,[6] and Troels C. Petersen[2]

[1]These authors contributed equally

[2]*Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen E*

[3]*Laboratoire de Physique, Ecole Normale Superieure, Rue Lhomond 15, Paris 07505*

[4]*Infektionsberedskab, Statens Serum Institute, Artillerivej, 2300 Copenhagen S*

[5]*DTU Compute, Section for Dynamical Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Anker Engelunds Vej 101A, 2800 Kongens Lyngby*

[6]*Animal Welfare and Disease Control, University of Copenhagen, Gronnegårdsvej 8, 1870 Frederiksberg C*

(Dated: October 9, 2020)

The modeling of Sars-CoV-2 has become a critical aspect of present life. In this fundamental task, a standard procedure has become the implementation of more or less detailed differential equations, which estimate the time development in the number of susceptible, exposed, infected, and recovered (SEIR) individuals. However, these SEIR models are based on very simple assumptions which constitute obvious approximations. In this paper, we introduce an agent based model that allows us to include spatial clustering effects and, based on Danish population data, we estimate how this impacts the long term development and the early prediction of a pandemic. Our results suggest that population clustering has a major impact on the long term development implying that our initial estimates on the herd immunity level, based on infection spreading in the early phases, might be over estimated by a factor of two.

Since the emergence of reports of a new disease from Wuhan, China [1], the causative pathogen now known as SARS-CoV-2 [2, 3] has spread dramatically, paralyzing societies, resulting in a large number of deaths and severe economic damage worldwide [4]. Mathematical models were developed to estimate the reproduction number and guide the authorities in an attempt to minimize the damage caused by this novel virus [5–9]. Generally, the models have been variants of the SIR model, and vary in complexity including simple deterministic compartmental models [9, 10], meta-population compartmental models [11–13], individual based models without including spatial specifications [6, 14, 15], and finally spatio-temporal agent-based models [16].

One important aspect in the modelling is to predict the herd immunity level, based on the early rise in the number of infected, before governmental interference. Earlier work has generally pointed out the importance of including heterogeneity when modeling the spread of infectious disease, as contact patterns between individuals [17], population mixing assumptions [18, 19] heterogeneities caused by super-spreaders [15] and the spatial dependency of COVID-19 [20, 21]. However, these mathematical models have not combined these elements or quantified how much, the early predictions for Sars-CoV-2 spread might be wrong.

* Electronic address: heltberg@nbi.ku.dk

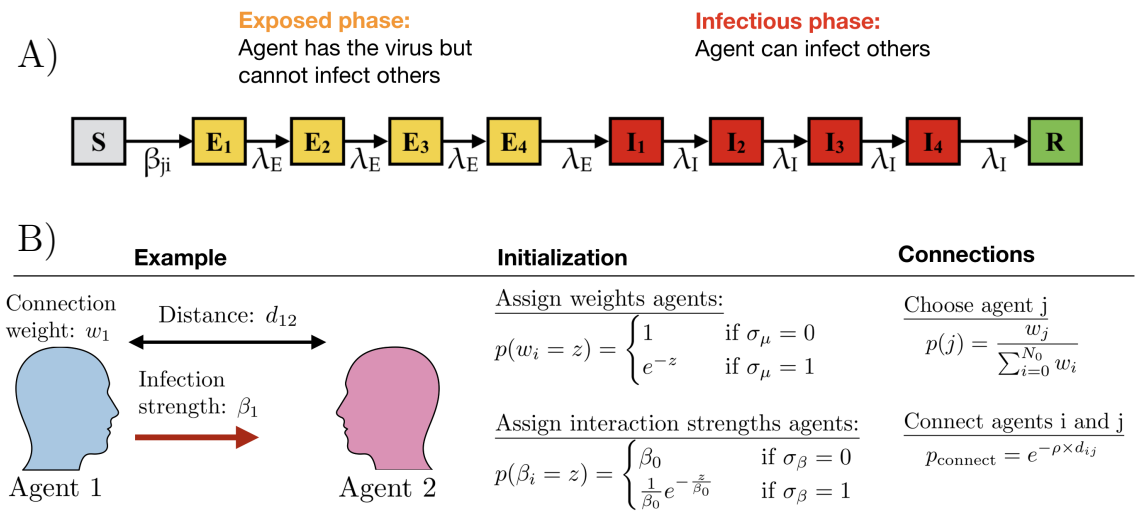

FIG. 1: **A)** Illustration of the modified susceptible-exposed-infected-removed (SEIR) model used. It consists of ten consecutive states (S, E_{1-4}, I_{1-4}, and R), with transition rates governed by β, λ_E, and λ_I, respectively. **B)** Illustration of how the spatial network is generated and heterogeneities in agents included.

We addressed these gaps by constructing an agent based model (ABM) on a realistic, spatially distributed population where each agent could be in one of four states and switch according to the rates schematized in Fig. 1A. We constructed a network of spatially distributed contacts (see S.I.) using data based on:

- The geographic location of people in Denmark (from Boligsiden [22])
- The average number of contacts per agent per day $\mu = 11$ (from HOPE [23])
- The average commuting distance $\rho = 0.1\,\mathrm{km}^{-1}$ (from Statistics Denmark [24])
- The fraction of long distance commutes $\epsilon_\rho = 4\,\%$ (from Statistics Denmark [24])

This is schematized in Fig. 1B, and all parameters in this model are found and explained in Table I. Heterogeneity in the distribution of connections in this network are created automatically, see Fig. 2A. We simulated the epidemic with 100 initially infected agents, $N_{\mathrm{init}} = 100$, and observed a spatial difference in areas affected by the disease (Fig. 2B). One region reached local herd-immunity (green arrow, Fig. 2B) while other regions of similar density were highly infected (red arrow, Fig. 2B) and yet other districts were almost unaffected (grey arrow, Fig. 2B). To quantify the effect of population clustering, we compared the ABM result to the SEIR model of similar parameters. This showed a significantly higher peak in the number of simultaneously infected agents, I_{max}, compared to the SEIR model and also that the epidemic happened faster (Fig. 2C). In contrary, when measuring the total number of agents who have been infected by the end of the epidemic, R_∞, we find that the ABM yielded a lower value than the SEIR model (Fig. 2D).

We investigated these discrepancies while varying ρ and found that spatial dependence leads to a significant rise in I_{max} for the ABM, $I_{\mathrm{max}}^{\mathrm{ABM}}$, compared to the SEIR model, $I_{\mathrm{max}}^{\mathrm{SEIR}}$, for $\beta = 0.01$ ($\mathcal{R}_0 \approx 2.1$ for $\rho = 0\,\mathrm{km}^{-1}$) and $\beta = 0.007$ ($\mathcal{R}_0 \approx 1.05$ for $\rho = 0\,\mathrm{km}^{-1}$) (black and blue points, Fig. 2E). Here $\mathcal{R}_0$ denotes the reproduction number. We introduced heterogeneity in infection strengths ($\sigma_\beta = 1$, see Fig. 1B) and found no significant effect (red points in Fig. 2E). Similarly, we introduced heterogeneity in connection weights ($\sigma_\mu = 1$, see Fig. 1B) which leads to a significant effect for $\rho = 0\,\mathrm{km}^{-1}$ but follows the other curves for $\rho > 0.1\,\mathrm{km}^{-1}$ (orange and green points in Fig. 2E). The total number of infected individuals when the epidemic is over, R_∞, decayed as a function of ρ except for $\beta = 0.007$ where the number of infected is larger than the SEIR model (Fig. 2F). Fixing $\rho = 0.1\,\mathrm{km}^{-1}$ and varying ϵ_ρ, we

Variable	Description	Value	Range	Units
N_0 :	Population size	$5.8 \cdot 10^6$	$10^5 - 10^7$	–
N_{init} :	Number of agents initially infected	100	$1 - 10^4$	–
μ :	Average number of network contacts	40	$10 - 100$	–
β :	Typical interaction strength	0.01	$0.001 - 0.1$	day^{-1}
λ_E :	Rate to move through ¼ of latency period	1	$0.5 - 4$	day^{-1}
λ_I :	Rate to move through ¼ of infectious period	1	$0.5 - 4$	day^{-1}
σ_μ :	Population clustering spread	0	$0 - 1$	–
σ_β :	Interaction strength spread	0	$0 - 1$	–
ρ :	Typical acceptance distance	0.1	$0 - 0.5$	km^{-1}
ϵ_ρ :	Fraction of distance-independent contacts	0.04	$0 - 1$	–

TABLE I: Overview of the seven parameters applied in this study, their typical value and the ranges we have considered. The first five parameters are standard SEIR parameters, whereas the last four parameters will induce specific spatial behaviour. These four parameters does not affect the SEIR model.

found that for $\epsilon_\rho < 0.5$ that the height $I_{\max}$ is almost unaffected (Fig. 2G). When ϵ_ρ increases, we found that R_∞^{ABM} increases linearly but never supersedes R_∞^{SEIR} (Fig. 2H).

Next we considered how these results could bias the predictions made from an the early rise in the number of infected (i.e. the curve to be flattened). These predictions were made from a fits on data from the early range, using a SEIR model (See Supplementary information). Without population clustering, the predicted curves fitted the number of infected individuals very well (Fig. 3A). Introducing population clustering ($\rho = 0.1\,\text{km}^{-1}$), leads to a severe overestimation of the disease based on the early curves (Fig. 3B). This result can be interpreted by the fact that in societies where population density and individual contact number is clustered, the early phase will be driven by people with many contacts, which typically happens in cities where population density is high. Varying the distance parameter ρ we found that the overestimation increases significantly even for very small spatial heterogeneities (Fig. 3C). This pattern turned out to be similar for the total number of infected predicted (Fig. 3D). We noted that the general trend for these results present for all our tested combinations of parameters and heterogeneities. To solidify this result, we varied ϵ_ρ and found that if long-distance connections are below 10 %, the size of the peak bias was constant within statistical uncertainty (Fig. 3E). For the total number of infected, we observed a decay, but even if $\epsilon_\rho = 0.25$, the discrepancy between prediction and result was still a factor of 2 (Fig. 3F). Taken together, these curves indicate the same thing; when one fits a SEIR model to infection numbers during the beginning of a pandemic, and use these estimates to predict the characteristics of the disease, one overestimates the effect of the disease by at least a factor of 2.

Our research reveals that the degree of population clustering in Denmark, creates a discrepancy between the early predictions made by the SEIR models and the underlying agent based interactions. This results in a significant overestimation of the impact of the disease, both in terms of maximal number of simultaneously infected (a factor of 3) and the total number of infected people (a factor of 2.5). Such discrepancies have been observed for earlier pandemics, for instance the 1918 Spanish flu, that the predicted herd immunity level was severely overestimated. These results can be an important element in explaining these mismatches, even though other elements, for instance social distancing and mutations to the viral strain, also play a part. During 2020, numerous countries have been faced with the task of laying out strategies to minimize the consequenses of SARS-CoV-2, including the importance of 'flattening the curve'. While this is truly crucial to avoid overpopulated hospitals, it should be taken seriously enough that we might specify to a higher degree of certainty which curve to be flattened. In the early phase, the

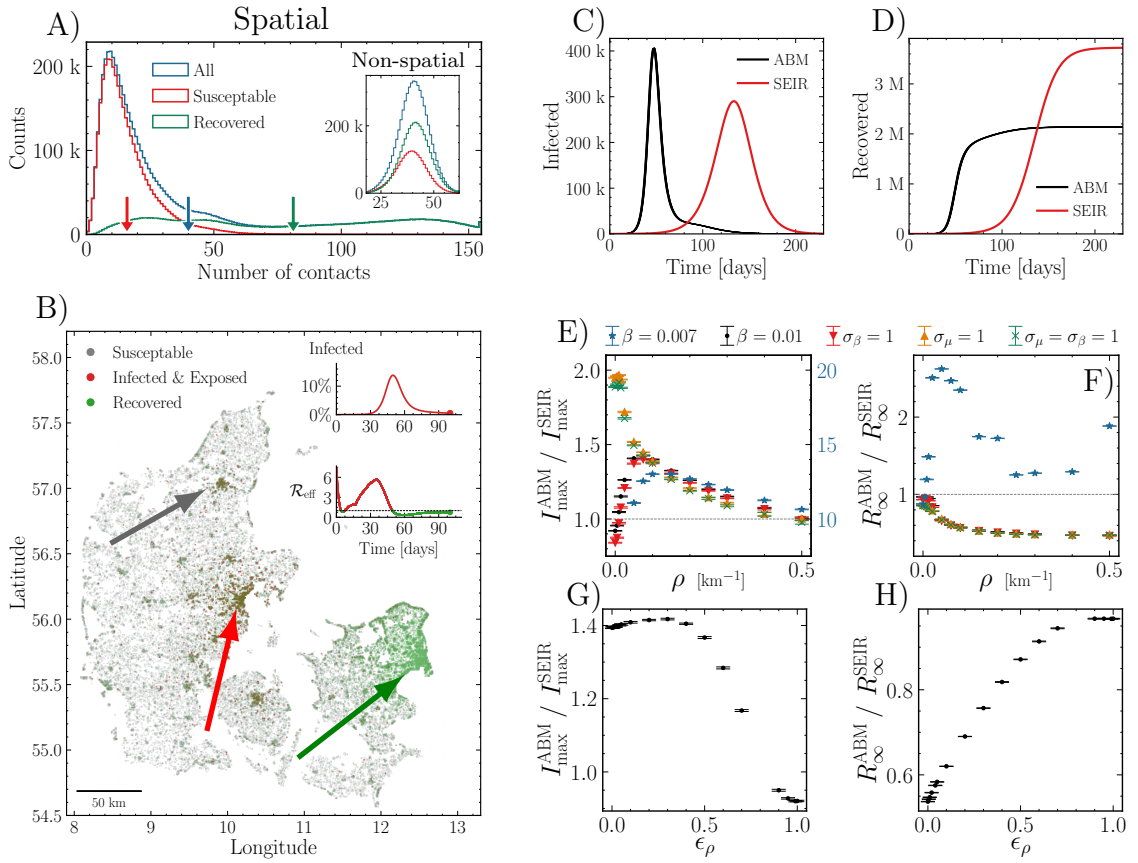

FIG. 2: **A)** Histograms showing the number of susceptible (red) and recovered (green) agents at the end of an epidemic with $\rho = 0.1$. The distribution before the epidemic is shown in blue. Arrows show the mean of each distribution. The inset shows the same for $\rho = 0$. **B)** Visualisation of the spatial position of infected agents during the infection. Green arrow: Largest city in Denmark (Copenhagen). Red Arrow: Second largest city in Denmark (Aarhus). Gray Arrow: Fourth largest city of Denmark (Aalborg). **C)** Number of infected agents as a function of time. Data shown for the spatially distributed network ($\rho = 0.1\,\mathrm{km}^{-1}$). Simulation was repeated 10 times. **D)** cumulative sum of agents who have had the disease as a function of time (with $\rho = 0.1\,\mathrm{km}^{-1}$). **E)** Relative difference in maximal number of infected, $I_{\max}$, between deterministic (SEIR) and ABM as a function of ρ, and shown for different parameters. Note the data for $\beta = 0.007$ are shown in blue with a factor 10 scaling (right y-axis). **F)** Relative difference in total number of infected at the end of the epidemic, R_∞, between deterministic (SEIR) and ABM as a function of ρ. Colors similar to **E**. **G)** Same as **E**, but as a function of ϵ_ρ. **H)** Same as **F**, but as a function of ϵ_ρ.

mean reproduction number across countries was around $2.5-3$ (see review [27]), leading to an estimation of 60% infected at the end of the pandemic. Our results estimate this number to be at least a factor of 2 too large. This has two important implications which are largely of a positive character. First, we do not have to fear as many infected as the well-mixed models predict. This could already huge impact on the outbreaks that will occur during the Autumn of 2020 and second waves could be smaller than expected, especially in countries that have already measured a high fraction of infected population. Secondly, this study emphasize the great benefits by making a lock-down early in spread of the pandemic. Since agents in the city-clusters are more likely to catch the infection, they are likely to be on average affected more in the beginning, and by removing contacts from these agents, one can avoid the worst peak while affecting the fewest number of people. During this pandemic, mathematical predictions have been heavily criticized ([25, 26]) and it is now important to improve the accuracy of these, in order to increase the confidence in the authorities, as models can be useful ([25]), as especially the early predictions will always lay the foundations for the politics of counter-measures and possibly lock-downs. This work seriously questions

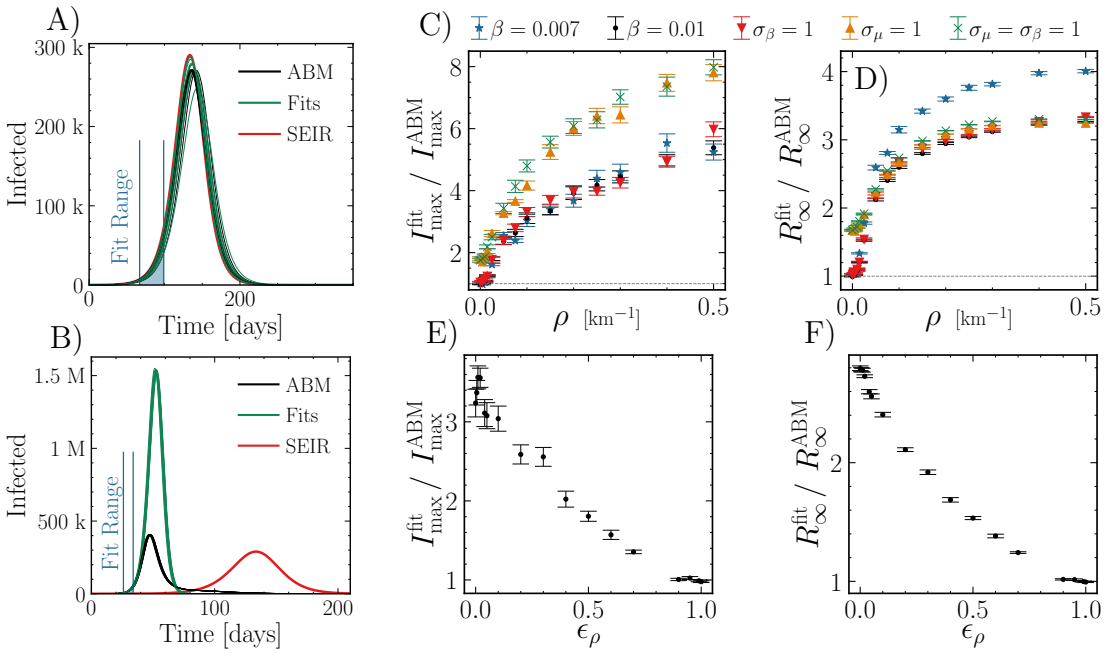

FIG. 3: **A)** Number of infected agents for the ABM in black, the SEIR model in red and the SEIR fits to the ABM data in green. Blue lines show the interval where parameters are fitted (also shown below the curves). Here $\rho = 0\,\text{km}^{-1}$. **B)** Same as **A** but with population clustering ($\rho = 0.1\,\text{km}^{-1}$). **C)** Relative difference in maximal number of infected, I_{max}, between the fit and the ABM for different values of ρ. Simulations repeated 10 times for each data-point. **D)** Relative difference in total number of infected at the end of the epidemic, R_∞, between the fit and the ABM for different values of ρ. **E)** Same as **C**, but as a function of ϵ_ρ. **F)** Same as **D**, but as a function of ϵ_ρ.

the validity of predictions using SEIR models and suggest that the precise effect of population clustering should be addressed more seriously for each country. While this work has absolutely no political agenda, it should serve as an input in the current debate of how to handle the severe consequences of a crisis like SARS-CoV-2.

Acknowledgements

The authors are grateful to the Danish expert group of SARS-CoV-2 modelling. Furthermore we thank Kim Sneppen for valuable discussions.

[1] Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... & Viboud, C. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science, 368(6489), 395-400.

[2] WHO: `www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19`. Accessed September 29, 2020.

[3] WHO: `www.who.int/csr/don/12-january-2020-novel-coronavirus-china`. Accessed September 29, 2020.

[4] Fernandes, N. (2020). Economic effects of coronavirus outbreak (COVID-19) on the world economy. Available at SSRN 3557504.

[5] Anderson, R. M., Heesterbeek, H., Klinkenberg, D.,& Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic?. The Lancet, 395(10228), 931-934.

[6] Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., ... & Flasche, S. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health.

[7] Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., ... & Dighe, A. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand.

[8] Keeling, M. J., Hollingsworth, T. D., & Read, J. M. (2020). The Efficacy of Contact Tracing for the Containment of the 2019 Novel Coronavirus (COVID-19). medRxiv.

[9] Kuniya, T. (2020). Prediction of the epidemic peak of coronavirus disease in Japan, 2020. Journal of clinical medicine, 9(3), 789.

[10] Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science, 368(6490), 489-493.

[11] Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., ... & Abbott, S. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health.

[12] van Bunnik, B. A., Morgan, A. L., Bessell, P., Calder-Gerver, G., Zhang, F., Haynes, S., ... & Lepper, H. C. (2020). Segmentation and shielding of the most vulnerable members of the population as elements of an exit strategy from COVID-19 lockdown. medRxiv.

[13] Danon, L., Brooks-Pollock, E., Bailey, M., & Keeling, M. J. (2020). A spatial model of COVID-19 transmission in England and Wales: early spread and peak timing. MedRxiv.

[14] Chang, S. L., Harding, N., Zachreson, C., Cliff, O. M., & Prokopenko, M. (2020). Modelling transmission and control of the COVID-19 pandemic in Australia. arXiv preprint arXiv:2003.10218.

[15] Sneppen, K., & Simonsen, L. (2020). Impact of Superspreaders on dissemination and mitigation of COVID-19. medRxiv.

[16] Milne, G. J., & Xie, S. (2020). The effectiveness of social distancing in mitigating COVID-19 spread: a modelling analysis. medRxiv.

[17] Bansal, S., Grenfell, B. T., & Meyers, L. A. (2007). When individual behaviour matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface, 4(16), 879-891.

[18] Kong, L., Wang, J., Han, W., & Cao, Z. (2016). Modeling heterogeneity in direct infectious disease transmission in a compartmental model. International journal of environmental research and public health, 13(3), 253.

[19] Brauer, F., Castillo-Chavez, C., & Castillo-Chavez, C. (2012). Mathematical models in population biology and epidemiology (Vol. 2, p. 508). New York: Springer.

[20] Kang, D., Choi, H., Kim, J. H., & Choi, J. (2020). Spatial epidemic dynamics of the COVID-19 outbreak in China. International Journal of Infectious Diseases.

[21] Giuliani, D., Dickson, M. M., Espa, G., & Santi, F. (2020). Modelling and predicting the spread of Coronavirus (COVID-19) infection in NUTS-3 Italian regions. arXiv preprint arXiv:2003.06664.

[22] Boligsiden: `www.boligsiden.dk`. Accessed September 29, 2020.

[23] HOPE project: `www.hope-project.dk`. Accessed September 29, 2020.

[24] Statistics Denmark: `www.statistikbanken.dk`. Accessed September 29, 2020.

[25] Holmdahl, I., & Buckee, C. (2020). Wrong but useful—what COVID-19 epidemiologic models can and cannot tell us. New England Journal of Medicine.

[26] Wynants, L., Van Calster, B., Bonten, M. M., Collins, G. S., Debray, T. P., De Vos, M., ... & Schuit, E. (2020). Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. bmj, 369.

[27] Boldog, P., Tekeli, T., Vizi, Z., Dénes, A., Bartha, F. A., & Röst, G. (2020). Risk assessment of novel coronavirus COVID-19 outbreaks outside China. Journal of clinical medicine, 9(2), 571