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 13, 2020)

The modeling of Sars-CoV-2 has become a critical aspect of present life. 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 (SIR/SEIR) individuals. However, these models are based on very simple assumptions which constitute obvious approximations. In this paper, we introduce an agent based model including spatial clustering. Based on Danish population data, we estimate how this impacts the early prediction of a pandemic and the long term developments. Our results show that early phase SEIR model predictions overestimate the peak number of infected and the equilibrium level by a at least a factor 2. These results are robust to variation of parameters influencing connection distances within a realistic range and independent of distribution of infection rates.

Since the emergence of reports of a new disease from Wuhan, China, the pathogen now known as SARS-CoV-2 has spread dramatically, paralyzing societies, resulting in a large number of deaths and severe economic damage worldwide [1–4]. 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/SEIR 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 the ability to predict the infection peak height and the steady state level of individuals having been infected based on the early rise in the number of infected before governmental interference. Earlier work has pointed out the importance of including heterogeneity when modeling the spread of infectious disease such 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 heterogeneous elements nor quantified how much the early SIR/SEIR predictions for Sars-CoV-2 might be biased.

* Electronic address: heltberg@nbi.ku.dk

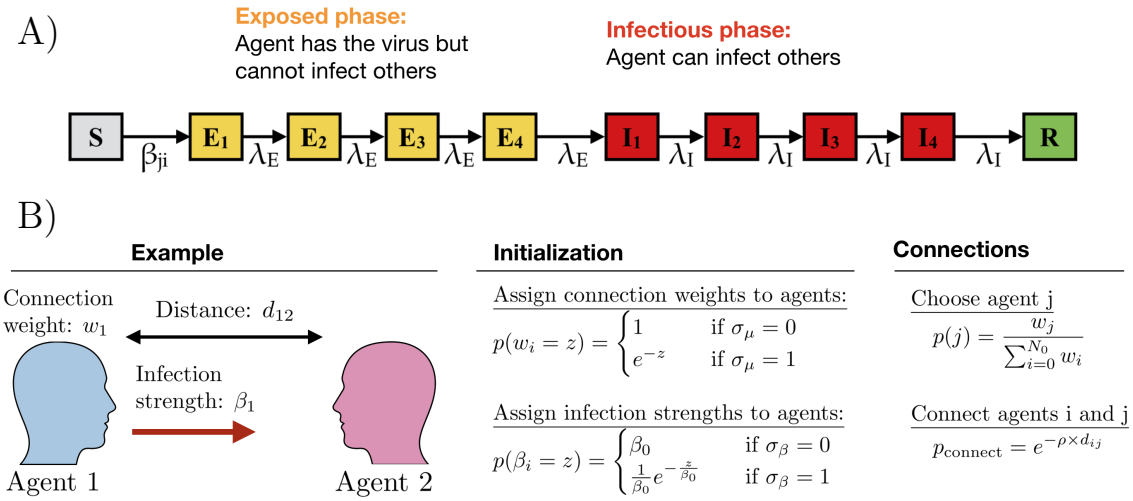

FIG. 1: **A)** Illustration of the modified susceptible-exposed-infected-removed (SEIR) model used. It consists of ten consecutive states (S, E_{1-4}, I_{1-4}, and R), with transition rates governed by β, λ_E, and λ_I, respectively. **B)** Illustration of how the spatial network is generated and heterogeneities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

We address these deficiencies by constructing an agent based model (ABM) on a realistic, spatially distributed population where each individual (agent) could be in one of four states and switch according to the rates schematized in Fig. 1A. We used this ABM to investigate the biases in SEIR models due to heterogeneities and in early prediction. We constructed a network of spatially distributed contacts (see Supplementary Information) using data based on:

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

This is schematized in Fig. 1B and all ten parameters in this model are explained in Table I. Heterogeneity in the distribution of connections in this network are created automatically through the population clustering, see Fig. 2A. This naturally leads to individuals living in densely populated areas having higher number of connections. We simulated the epidemic with 100 initially infected individuals, $N_{\mathrm{init}} = 100$, and observed a spatial difference in areas affected by the disease (Fig. 2B). One region reached local endemic steady state (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 reference SEIR model of similar parameters. This showed a significantly higher peak in the number of simultaneously infected individuals, I_{peak}, compared to the SEIR model and also that the epidemic developed at a higher rate (Fig. 2C). In contrary, when measuring the total number of individuals who have been infected by the end of the epidemic, R_∞, we find that the ABM yielded a lower number than the SEIR model (Fig. 2D).

For reference, we chose two values of infection rates, $\beta = 0.01$ and 0.007, which in the absence of spatial dependence (i.e. $\rho = 0\,\mathrm{km}^{-1}$) corresponds to initial reproduction numbers $\mathcal{R}_0 \approx 2.1$ and 1.05, respectively. Varying ρ, we found that spatial dependence leads to a significant rise in I_{peak} for the ABM, $I^{\mathrm{ABM}}_{\mathrm{peak}}$, compared to the SEIR model, $I^{\mathrm{SEIR}}_{\mathrm{peak}}$ (black and blue points, Fig. 2E). We introduced heterogeneity in infection strengths ($\sigma_\beta = 1$, see Fig. 1B), thus introducing a distribution in infection strengths (i.e. including '*super shedders*'). We found no significant impact from this effect (red points in Fig. 2E). Similarly, we introduced heterogeneity in connection weights ($\sigma_\mu = 1$, see Fig. 1B), thus introducing a distribution in number of connections (i.e. including '*super connecters*'). This leads to a significant effect

Variable	Description	Value	Range	Units
N_0 :	Population size	$5.8 \cdot 10^6$	$10^5 - 10^7$	–
N_{init} :	Number of individuals initially infected	100	$1 - 10^4$	–
μ :	Average number of network contacts	40	$10 - 100$	–
β :	Typical infection 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 ten parameters applied in this study, their typical value, and the ranges we have considered. The first six parameters are standard SEIR parameters, whereas the last four parameters will induce specific spatial behaviour. These four parameters do not affect the SEIR model.

for $\rho = 0\,\text{km}^{-1}$ but follows the other curves for $\rho > 0.1\,\text{km}^{-1}$ (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 endemic steady state level is larger than the one obtained by the SEIR model (Fig. 2F). Fixing $\rho = 0.1\,\text{km}^{-1}$ and varying the fraction of distance-independent contacts, ϵ_ρ, we found that for $\epsilon_\rho < 0.5$ the peak I_{peak} 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 consider how these heterogeneities bias the traditional SEIR predictions, especially the predictions based on fits of the number of infected (i.e. the curve to be flattened) in the beginning of the epidemic (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 infections (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 the 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 predictions of the total number of infected (Fig. 3D). We observed this general trend for all tested combinations of parameters and heterogeneities. To solidify our results, we varied ϵ_ρ and found that if long-distance connections are below 10 %, the bias in the estimation of the infection peak height, I_{peak}, was constant within statistical uncertainty (Fig. 3E). For the total number of infected, R_∞, we observed an almost linear regression to the SEIR model. However, even when $\epsilon_\rho = 0.25$, the prediction bias was still a factor of 2 (Fig. 3F). Taken together, these curves indicate a general trend; if 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. It results in a significant overestimation of the impact of the disease, both in terms of maximal number of simultaneously infected (by a factor of 3) and the total number of infected people (by a factor of 2.5). Such discrepancies have been observed for earlier pandemics, for instance the 1918 Spanish flu, in the sense that the predicted herd immunity level was severely overestimated [25]. These results can be an important element in explaining these mismatches, even though other elements, as 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 consequences of SARS-CoV-2, including the importance of '*flattening*

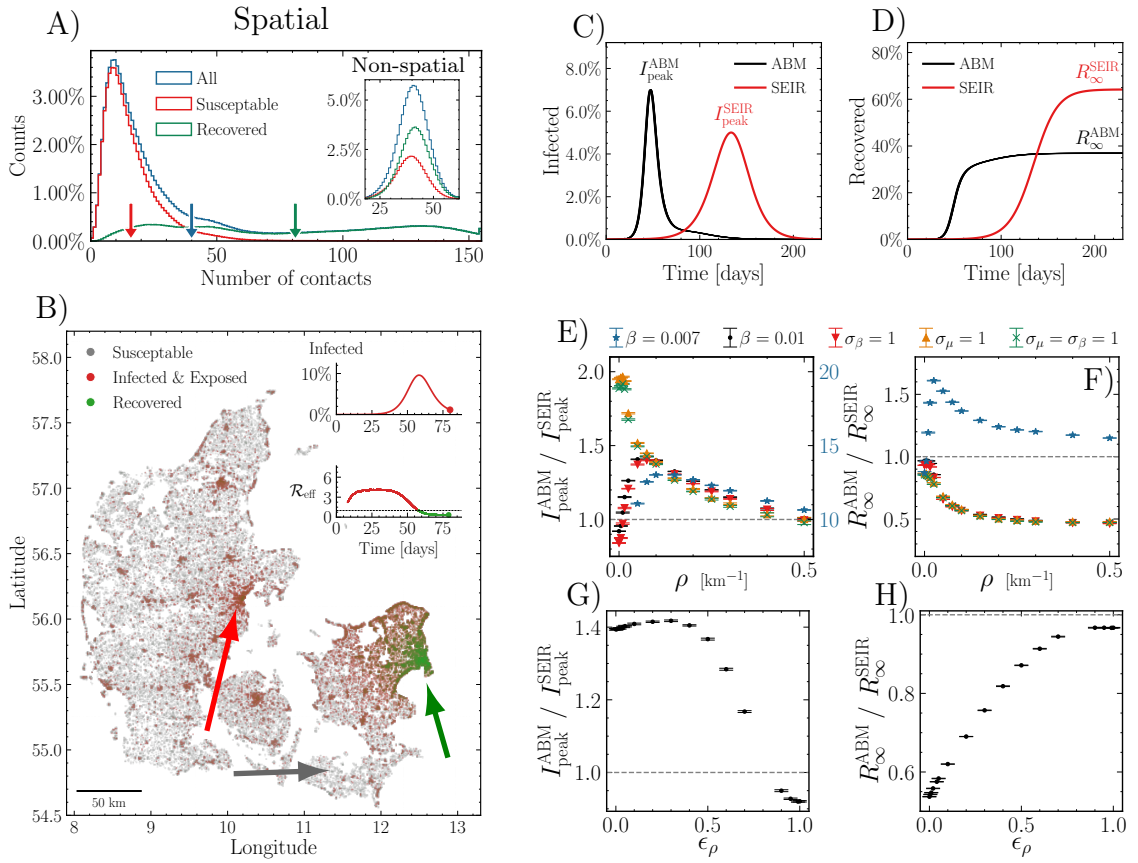

FIG. 2: **A)** Histograms showing the number of susceptible (red) and recovered (green) individuals at the end of an epidemic with $\rho = 0.1\,\mathrm{km}^{-1}$. The distribution before the epidemic is shown in blue. The arrows show the mean of each distribution. The inset shows the same for $\rho = 0\,\mathrm{km}^{-1}$. **B)** Visualisation of the spatial position of individuals during the infection and which state they are in. Green arrow: Largest city in Denmark (Copenhagen): mostly recovered. Red arrow: Second largest city in Denmark (Aarhus): mostly infected. Grey arrow: low-population area: mostly susceptible (i.e. have not been infected). **C)** Number of infected individuals 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 individuals 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_{peak}, 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.

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. The mean reproduction number, $\mathcal{R}_0$, was around $2.5 - 3$ across countries in the early phase [28]. This lead to estimations of predicting that 60% of the populations would be 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 have 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 in the population. Secondly, our study emphasize the great benefits by making lock-downs early in the pandemic before the disease starts to spread too much. Since people in living in city-clusters are more likely to catch the infection, they are, on average, more likely to be affected in the beginning and by removing contacts from these individuals one can avoid the worst peak while affecting

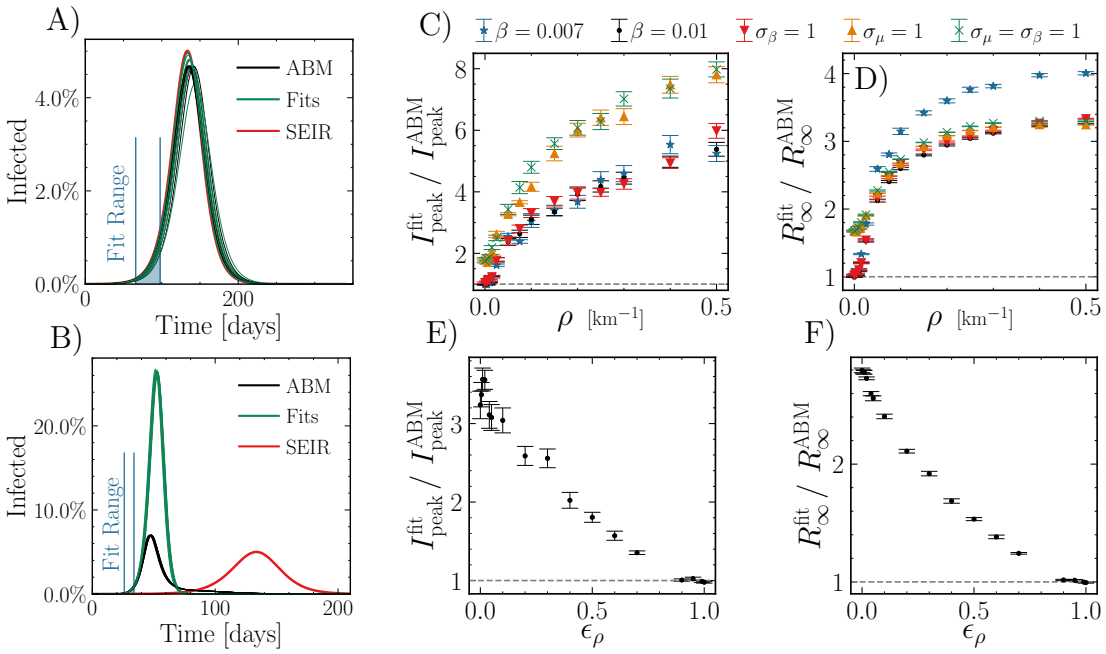

FIG. 3: **A)** Number of infected individuals for the ABM in black, the SEIR model in red and the SEIR fits to the ABM data in green. Blue lines show the interval where parameters are fitted (also shown below the curves). Here $\rho = 0\,\text{km}^{-1}$. **B)** Same as **A** but with population clustering ($\rho = 0.1\,\text{km}^{-1}$). **C)** Relative difference in maximal number of infected, I_{peak}, 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 fewest number of people. During this pandemic, mathematical predictions have been heavily criticized [26, 27] and it is now important to improve the accuracy of them, in order to increase the confidence in the authorities. Models can be useful [26], especially as the early predictions will lay the foundations for the political initiatives such as counter-measures and lock-downs. Our work seriously questions the validity of predictions based on SEIR models and suggests that the precise effect of population clustering should be addressed more seriously for each country. While our 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., et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College London (16-03-2020), doi: www.doi.org/10.25561/77482.

[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., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Jit, M., & Klepac, P. (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, 5(5), e261–e270. www.doi.org/10.1016/S2468-2667(20)30073-6.

[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 2020.05.04.20090597; doi: www.doi.org/10.1101/2020.05.04.20090597.

[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 2020.02.12.20022566; www.doi.org/10.1101/2020.02.12.20022566.

[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 2020.05.17.20104745; www.doi.org/10.1101/2020.05.17.20104745.

[16] Milne, G. J., & Xie, S. (2020). The effectiveness of social distancing in mitigating COVID-19 spread: a modelling analysis. medRxiv 2020.03.20.20040055; www.doi.org/10.1101/2020.03.20.20040055.

[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. www.doi.org/10.1016/j.ijid.2020.03.076.

[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] Andreasen, V., Viboud, C., & Simonsen, L. (2008). Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies. The Journal of infectious diseases, 197(2), 270-278.

[26] Holmdahl, I., & Buckee, C. (2020). Wrong but useful—what COVID-19 epidemiologic models can and cannot tell us. New England Journal of Medicine. www.doi.org/10.1056/NEJMp2016822.

[27] 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.

[28] 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