Inferring between-population differences in COVID-19 dynamics

Bryan Wilder[1*], **Marie Charpignon**[2], **Jackson A. Killian**[1], **Han-Ching Ou**[1], **Aditya Mate**[1], **Shahin Jabbari**[1], **Andrew Perrault**[1], **Angel Desai**[3], **Milind Tambe**[1*], **Maimuna S. Majumder**[4,5*]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[2]MIT Institute for Data, Systems, and Society, Cambridge, MA, USA
[3]International Society for Infectious Diseases, Brookline, MA, USA
[4]Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[5]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA

Abstract

As the COVID-19 pandemic continues, formulating targeted policy interventions supported by differential SARS-CoV2 transmission dynamics will be of vital importance to national and regional governments. We develop an individual-level model for SARS-CoV2 transmission that accounts for location-dependent distributions of age, household structure, and comorbidities. We use these distributions together with age-stratified contact matrices to instantiate specific models for Hubei, China; Lombardy, Italy; and New York, United States. We then develop a Bayesian inference framework which leverages data on reported deaths to obtain a posterior distribution over unknown parameters and infer differences in the progression of the epidemic in the three locations. These findings highlight the role of between-population variation in formulating policy interventions.

Introduction

Since December 2019, the COVID-19 pandemic – caused by the novel coronavirus SARS-CoV2 – has resulted in significant morbidity and mortality [Baud *et al.*, 2020]. As of May 19, 2020, an estimated 4,800,000 individuals have been infected, with over 318,000 fatalities worldwide [Center for Systems Science and Engineering at Johns Hopkins University, 2020]. Key factors such as existing comorbidities and age have appeared to play a role in an increased risk of mortality [Zhou *et al.*, 2020]. Epidemiological studies have provided significant insights into the disease to date [Xu *et al.*, 2020; Riou and Althaus, 2020; Li *et al.*, 2020; Kucharski *et al.*, 2020]. However, as national and regional governments begin to implement broad-reaching policies in response to rising case counts and stressed healthcare systems, tailoring these polices based on an understanding of how population-specific demography impacts outbreak dynamics will be vital. Previous modeling studies have largely not incorporated the rich set of household demographic features needed to address such questions.

This study employs mathematical modeling to assess how the distribution of age, comorbidities, and household contacts in a population impact the utility of potential non-pharmaceutical interventions and overall transmission dynamics. We develop a stochastic agent-based model for SARS-CoV2 transmission which accounts for distributions of age, household types, comorbidities, and contact between different age groups in a given population (Fig. 1). Our model accounts for both within-household contact (simulated via household distributions taken from census data) and out-of-household contact using age-stratified, country-specific estimated contact matrices [Prem *et al.*, 2017]. We instantiate the model for Hubei, China; Lombardy, Italy; and New York, United States, developing a Bayesian inference strategy for estimating the distribution of unknown parameters using data on reported deaths in each location. This enables us to tease out differences in the progression of the epidemic in the three locations.

Model description

We develop an agent-based model for COVID-19 spread which accounts for the distributions of age, household types, comorbidities, and contact between different age groups in a given population. The model follows a *susceptible-exposed-infectious-removed (SEIR)* template [Van den Driessche *et al.*, 1999; Ball *et al.*, 2015]. Specifically, we simulate a population of n agents (or individuals), each with an age a_i, a set of comorbidities c_i, and a household (a set of other agents). We stratify age into ten-year intervals and incorporate hypertension and diabetes as comorbidities due to their worldwide prevalence [Roth *et al.*, 2018] and association with higher risk of in-hospital death for COVID-19 patients [Zhou *et al.*, 2020]. We track agents through the process of social contact, becoming infected with the disease, and progressing through more severe forms of the disease until either death or recovery.

The disease is transmitted over a contact structure, which is divided into in-household and out-of-household groups. Each agent has a household consisting of a set of other agents (see the SI for details on how households are generated using country-specific census information). Individuals infect members of their households at a higher rate than out-of-household agents. We model out-of-household transmission

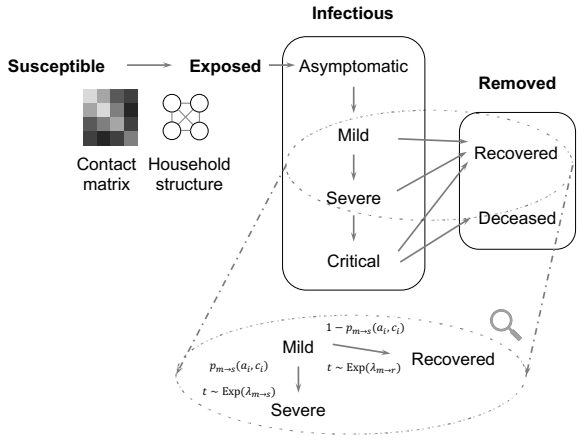

Figure 1: We use a modified SEIR model, where the infectious states are subdivided into levels of disease severity. The transitions are probabilistic and there is a time lag for transitioning between states. For example, the magnified section shows the details of transitions between mild, recovered, and severe states. Each arrow consists of the probability of transition (e.g., $p_{m\to s}(a_i, c_i)$ denotes the probability of progressing from mild to severe) as well as the associated time lag (e.g., the time t for progression from mild to severe is drawn from an exponential distribution with mean $\lambda_{m\to s}$). a_i and c_i denote the age and set of comorbidities of the infected individual i.

using country-specific estimated contact matrices [Prem *et al.*, 2017]. These matrices state the mean number of daily contacts an individual of a particular age strata has with individuals from each of the other age strata.

The model iterates over a series of discrete time steps, each representing a single day, from a starting time t_0 to an end time T. There are two main components to each time step: disease progression and new infections. The progression component is modeled by drawing two random variables for each individual each time they change severity levels (e.g. on entering the mild state). The first random variable is Bernoulli and indicates whether the individual will recover or progress to the next severity level. The second variable represents the amount of time until progression to the next severity level. We use exponential distributions for almost all time-to-event distributions, a common choice in the absence of specific distributional information [Allison, 2010; Collett, 2015]. The exception is the incubation time between asymptomatic and mild states, where more specific information is available; here, we use a log-normal distribution based on estimates by [Lauer *et al.*, 2020]. Details on parameter choices, including estimating the probability of progression by age and comorbidity status, are deferred to a full version of the paper.

In the new infections component, infected individuals infect each of their household members with probability p_h at each time step. p_h is calibrated so that the total probability of infecting a household member before either isolation or recovery matches the estimated secondary attack rate for household members of COVID-19 patients (i.e., the average fraction of household members infected) [Liu *et al.*, 2020]. Infected individuals draw outside-of-household contacts from the general population using the country-specific contact matrix. For an infected individual of age group i, we sample $w_{ij}^s \sim \text{Poisson}(M_{ij}^s)$ contacts for each age group j and setting s where M^s is the country-specific contact matrix for setting s (from [Prem *et al.*, 2017]). We include contacts in work, school, and community settings. Then, we sample w_{ij}^s contacts of age j uniformly with replacement, and each contact is infected with the probability p_{inf}, the probability of infection given contact.

Inference of posterior distributions

We infer unknown model parameters and states in a Bayesian framework. This entails placing a prior distribution over the unknown parameters, and then specifying a likelihood function for the observable data, the time series of deaths reported in a location. We posit the following generative model for the observed deaths:

$$
\begin{aligned}
&p_{\text{inf}}, d_{\text{mult}}, t_0 \sim \mathcal{U} \\
&d_1 .... d_T \sim ABM(p_{\text{inf}}, d_{\text{mult}}, t_0) \\
&o_t \sim \text{NegativeBinomial}(d_t, \sigma_{obs}^2)\ t = 1...T
\end{aligned}
$$

where $\mathcal{U}$ denotes a joint uniform prior, ABM denotes a draw from the stochastic agent-based dynamics, $d_1 ... d_T$ are the time series output by the simulation, and $o_1 ... o_T$ are the number of deaths observed on the corresponding dates. We model the observations as drawn from a negative binomial distribution (appropriate for overdispersed count data) with dispersion parameter σ_{obs}^2. We separately estimated σ_{obs}^2 by fitting an autoregressive negative binomial regression to the observed counts using the R package tscount [Liboschik *et al.*, 2015]. The negative binomial observation model was strongly preferred to a Poisson model by AIC values. Together, the likelihood function is given by

$$
\mathcal{L}(p_{\text{inf}}, d_{\text{mult}}, t_0, d_1 ... d_T) = \prod_{t=1}^{t} \Pr\left[o_t | d_t, \sigma_o^2\right].
$$

To obtain the posterior distribution, we use Latin hypercube sampling to draw many (10-80 thousand per location, depending on the size of the prior ranges) samples from the joint uniform prior over $p_{\text{inf}}, d_{\text{mult}}$ and t_0, and then sample the latent variables $d_1 ... d_T$ at each combination of parameters. We compute the likelihood for the full sample (including the latent variables). This allows us to use importance sampling to resample values of $(p_{\text{inf}}, d_{\text{mult}}, t_0, d_1 ... d_T)$ according to the posterior distribution. Finally, we marginalize out $d_1 ... d_T$ to obtain the posterior over the parameters $p_{\text{inf}}, d_{\text{mult}}, t_0$, along with unobservable state variables of the simulation such as the number of infected individuals at each step.

Application to Hubei, Lombardy, and New York City

Using the model, we estimate posterior distributions over unobserved quantities which characterize the dynamics of the epidemic in a particular location. We present estimates for two quantities. First, the basic reproduction number r_0. Second, the rate at which infections are documented. Neither

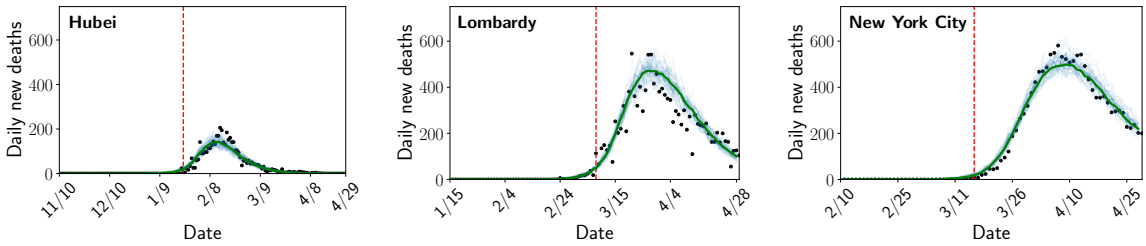

Figure 2: Posterior distribution over the number of deaths each day compared to the true number of reported deaths. Light blue lines are individual samples from the posterior, green is the median, and the black dots are the number of reported deaths. The red dashed line represents the start of modeled contact reductions in each location. Left to right: Hubei, Lombardy, New York.

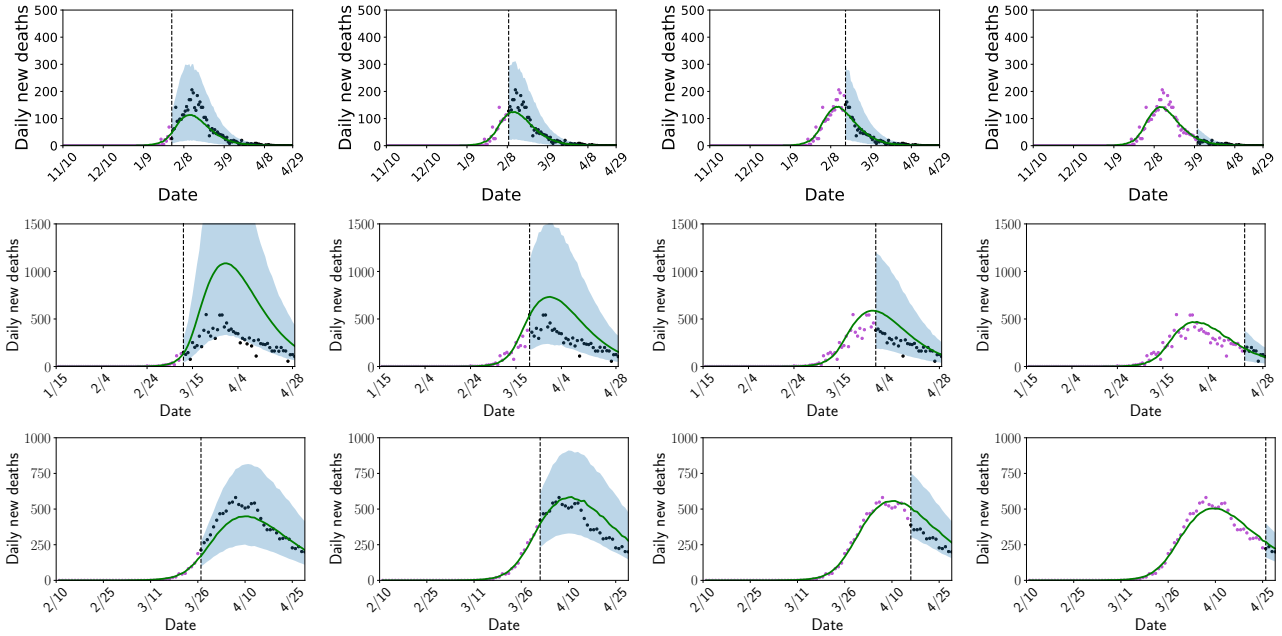

Figure 3: Predictive posterior for each location as a function of when the training period ends. Top row: Hubei; middle row: Lombardy; bottom row: New York. Black dashed line: end of training period. Green line: posterior median. Blue shaded region: 90% credible interval. Pink dots: training data. Black dots: held-out data. The 90% credible interval of the predictive posterior includes contains the held-out data at nearly all points, including when the model is fit using only data from the earliest portion of the epidemic.

quantity is directly observable in the data due to substantial underdocumentation of infections. However, these estimates are needed to characterize the scope of the outbreak in a particular place, the degree to which existing testing strategies capture new infections, and the rate at which infections are expected to increase in the absence of any intervention. These findings are critical to formulate policy interventions which are responsive to the outbreak as it evolves in a given population.

There are three parameters for which values are not precisely estimated in the literature, and which we place prior distributions over. First is p_{inf}, the probability of infection given contact with an infected individual. This determines the level of transmissibility of the disease. Second, t_0, the start time of the infection, which is not precisely characterized in most locations and has an impact due to rapid doubling times. Third, a parameter d_{mult}, which accounts for differences in mortality rates between locations that are *not* captured by demographic factors in the model (e.g., the impact of variation in health system capacities). d_{mult} is a multiplier applied to the baseline mortality rate from [Verity *et al.*, 2020]. We incorporate reduced person-to-person contact after mobility restrictions were imposed in each location, basing the strength of the effect on post-lockdown contact surveys [Zhang *et al.*, 2020] or mobility data from mobile phones [Google, 2020].

By conditioning on the observed time series of deaths, we obtain a joint posterior distribution over both the three unknown parameters and unobserved model states such as the number of people infected at each time step. We use reported deaths because they are believed to be better documented than infections, and perform a sensitivity analysis to account for possible underdocumentation of deaths [Katz *et al.*, 2020; Modi *et al.*, 2020].

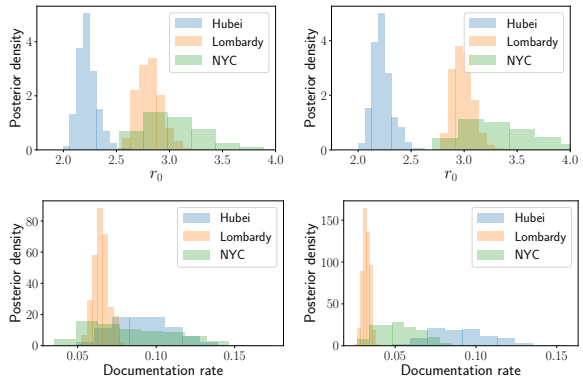

Figure 4: Posterior distribution over r_0 and the fraction of infections documented in each location. Left: conditioning on reported deaths. Right: conditioning on deaths being twice what was reported.

Validation

Fig. 2 shows that the posterior distribution of the model closely reproduces the observed time series of deaths in each location. As an additional check, Figure 3 shows validation on out-of-sample data for each location. Each plot shows the result of conditioning on observations only up until a specified time point. Then, we show the predictive posterior distribution over the data which was not used for training. Even when given training data only from the early stages of the epidemic, the model is able to capture the basic features of the outbreak. The timing of the peak in deaths is matched well across all three locations, and the observed deaths nearly always lie within the 90% credible interval of the predictive posterior. As expected, the fit improves as more training data becomes available. However, the model's predictions become substantially more accurate even before the peak in deaths is observed in the training data (the second column). While accurate forecasting is not our primary aim (rather, our goal is to make inferences about the dynamics of the outbreak as it occurred in each location), reasonable behavior by the predictive posterior helps guard against the possibility of overfitting.

Inferring differences in dynamics between populations

The top row of Fig. 4 shows the posterior distribution over r_0 in each location. Substantial differences are evident between the three locations. The posterior median is 2.21 in Hubei (90% credible interval: 2.10–2.41), 2.80 in Lombardy (2.66–3.01), and 3.06 in New York (2.65–3.59). The estimates for Hubei fall within the range of a number of existing estimates [Majumder and Mandl, 2020], while the interval for Lombardy is lower than the interval 2.9–3.2 estimated by previous work [Guzzetta *et al.*, 2020]. The estimated r_0 for New York is larger than either Hubei or Lombardy. The main between-population differences are not impacted by a sensitivity analysis for underreporting of deaths, shown in Figure 4. Death totals from Hubei have been substantially revised upwards to correct for underreporting in the early stages of the epidemic [British Broadcasting Corporation, 2020], but such corrections are either unavailable or rapidly evolving for Lombardy and New York. Our sensitivity analysis assumes that deaths in Lombardy and New York are twice what was reported, consistent with excess mortality data [Katz *et al.*, 2020; Modi *et al.*, 2020]. The comparison across populations is unaltered.

The bottom row of Fig. 4 shows the posterior distribution over the fraction of infections which were documented in each location (obtained by dividing the number of confirmed cases in each location by the number of infections in the simulation under each sample from the posterior). Documentation rates are uniformly low, indicating undocumented infections in all locations. However, we estimate lower documentation in Lombardy (90% credible interval: 5.7–7.3%) than in either New York (4.8–13.1%) or Hubei (6.5–12.2%). While this general trend is consistent with previous estimates by Russell et al. [Russell *et al.*, 2020], our estimates are substantially lower. One potential explanation is that Russell et al. estimate documentation from death data using a case fatality rate (CFR) from the literature while our model uses a *infection* fatality rate (IFR). The IFR is lower because it includes all infections, not only those that become confirmed cases. This approach requires more infections to account for a given number of deaths.

Although we estimate more undocumented infections, all locations remain potentially vulnerable to second-wave outbreaks, with the median percentage of the population infected at 7.5% in Hubei, 11.7% in Lombardy and 25.2% in New York. Recent serological surveys have estimated 25% infected in New York [CBS New York, 2020], consistent with our distribution. When assuming twice the number of reported deaths, the median percentage infected is 23.4% in Italy and 38.6% in New York. Overall, our estimates for r_0 and the remaining population of susceptible individuals indicate that Hubei, Lombardy, and New York could experience new outbreaks in the absence of continued interventions to reduce transmission. Despite this, between-population differences remain substantial: Hubei, Lombardy, and New York have had distinct experiences with COVID-19 which will continue to shape future policy responses.

Conclusion

We developed an agent-based model of SARS-COV2 transmission which accounts for population-specific demographic structure, along with a Bayesian framework for conducting inference of unknown model parameters. We observe wide variation across locations in the dynamics and progression of the epidemic. Ongoing work will explore the policy implications of these results, tracing out how the impact of interventions may differ between populations. We hope that developing methods which allow a finer-grained understanding of the COVID-19 epidemic will help policymakers formulate more targeted and effective responses.

Acknowledgments

This work was supported in part by the Army Research Office by grant MURI W911NF1810208 and in part by grant T32HD040128 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health. Killian was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303. Perrault and Jabbari were supported by the Harvard Center for Research on Computation and Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[Allison, 2010] Paul Allison. *Survival analysis using SAS: a practical guide*. SAS Institute, 2010.

[Ball *et al.*, 2015] Frank Ball, Edward Knock, and Philip O'Neill. Stochastic epidemic models featuring contact tracing with delays. *Mathematical Biosciences*, 266:23–35, 2015.

[Baud *et al.*, 2020] David Baud, Xiaolong Qi, Karin Nielsen-Saines, Didier Musso, Léo Pomar, and Guillaume Favre. Real estimates of mortality following COVID-19 infection. *The Lancet*, 2020.

[British Broadcasting Corporation, 2020] British Broadcasting Corporation. Coronavirus: China outbreak city Wuhan raises death toll by 50%, 2020. https://www.bbc.com/news/world-asia-china-52321529, Last Accessed: 2020-05-17.

[CBS New York, 2020] CBS New York. Coronavirus Antibodies Present In Nearly 25% Of All NYC Residents, 2020. https://newyork.cbslocal.com/2020/04/27/coronavirus-antibodies-present-in-nearly-25-of-all-nyc-residents/, Last Accessed: 2020-05-17.

[Center for Systems Science and Engineering at Johns Hopkins University, 2020] Center for Systems Science and Engineering at Johns Hopkins University. Coronavirus COVID-19 Global Cases, 2020. https://coronavirus.jhu.edu/map.html.

[Collett, 2015] David Collett. *Modelling survival data in medical research*. CRC Press, 2015.

[Google, 2020] Google. COVID-19 community mobility reports, 2020. https://www.google.com/covid19/mobility/.

[Guzzetta *et al.*, 2020] Giorgio Guzzetta, Piero Poletti, Marco Ajelli, Filippo Trentini, Valentina Marziano, Danilo Cereda, Marcello Tirani, Giulio Diurno, Annalisa Bodina, Antonio Barone, Lucia Crottogini, Maria Gramegna, Alessia Melegaro, and Stefano Merler. Potential short-term outcome of an uncontrolled COVID-19 epidemic in Lombardy, Italy, February to March 2020. *Eurosurveillance*, 25(12), 2020.

[Katz *et al.*, 2020] Josh Katz, Denise Lu, and Margot Sanger-Katz. What is the real coronavirus death toll in each state? *The New York Times*, 2020. https://www.nytimes.com/interactive/2020/05/05/us/coronavirus-death-toll-us.html.

[Kucharski *et al.*, 2020] Adam Kucharski, Timothy Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, and Rosalind Eggo. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 2020.

[Lauer *et al.*, 2020] Stephen Lauer, Kyra Grantz, Qifang Bi, Forrest Jones, Qulu Zheng, Hannah Meredith, Andrew Azman, Nicholas Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 2020.

[Li *et al.*, 2020] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*, 2020.

[Liboschik *et al.*, 2015] Tobias Liboschik, Konstantinos Fokianos, and Roland Fried. tscount: An r package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, 2015.

[Liu *et al.*, 2020] Yang Liu, Rosalind Eggo, and Adam Kucharski. Secondary attack rate and superspreading events for SARS-CoV-2. *The Lancet*, 2020.

[Majumder and Mandl, 2020] Maimuna Majumder and Kenneth Mandl. Early in the epidemic: Impact of preprints on global discourse of 2019-nCOV transmissibility. *SSRN*, 2020.

[Modi *et al.*, 2020] Chirag Modi, Vanessa Boehm, Simone Ferraro, George Stein, and Uros Seljak. Total covid-19 mortality in italy: Excess mortality and age dependence through time-series analysis. *medRxiv*, 2020.

[Prem *et al.*, 2017] Kiesha Prem, Alex Cook, and Mark Jit. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS Computational Biology*, 13(9):e1005697, 2017.

[Riou and Althaus, 2020] Julien Riou and Christian Althaus. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance*, 25(4), 2020.

[Roth *et al.*, 2018] Gregory Roth, Degu Abate, Kalkidan Hassen Abate, Solomon Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159):1736–1788, 2018.

[Russell *et al.*, 2020] Timothy Russell, Joel Hellewell, Sam Abbott, Christopher Jarvis, Kevin van Zandvoort, Stefan Flasche, Rosalind Eggo, John Edmunds, and Adam Kucharski. Using a delay-adjusted case fatality ratio to estimate under-reporting, 2020. https://cmmid.github.io/topics/covid19/severity/global_cfr_estimates.html. Accessed 03-26-20.

[Van den Driessche *et al.*, 1999] Pauline Van den Driessche, Michael Li, and James Muldowney. Global stability of SEIRS models in epidemiology. *Canadian Applied Mathematics Quarterly*, 7:409–425, 1999.

[Verity *et al.*, 2020] Robert Verity, Lucy Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick Walker, Han Fu, et al. Estimates of the severity of COVID-19 disease. *medRxiv*, 2020.

[Xu *et al.*, 2020] Bo Xu, Alomía de Gutiérrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily Cohn, Yulin Hswen, Sarah Hill, María Mercedes Cobo, Alexander Zarebski, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific Data*, 7, 2020.

[Zhang *et al.*, 2020] Juanjuan Zhang, Maria Litvinova, Yuxia Liang, Yan Wang, Wei Wang, Shanlu Zhao, Qianhui Wu, Stefano Merler, Cecile Viboud, Alessandro Vespignani, et al. Changes in contact patterns shape the dynamics of the covid-19 outbreak in china. *Science*, 2020.

[Zhou *et al.*, 2020] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 2020.