

A preceding low-virulence strain pandemic inducing immunity against COVID-19

Hagai B. Perets^{1*} and Ruth Perets^{1,2}

July 12, 2020

¹Technion - Israel Institute of Technology, Haifa, Israel 3200004

²Rambam Health Care Campus, Haifa, Israel

Abstract

The COVID-19 pandemic is thought to began in Wuhan, China in December 2019. Mobility analysis identified East-Asia and Oceania countries to be highly-exposed to COVID-19 spread, consistent with the earliest spread occurring in these regions. However, here we show that while a strong positive correlation between case-numbers and exposure level could be seen early-on as expected, at later times the infection-level is found to be *negatively* correlated with exposure-level. Moreover, the infection level is positively correlated with the population size, which is puzzling since it has not reached the level necessary for population-size to affect infection-level through herd immunity. These issues are resolved if a low-virulence Corona-strain (LVS) began spreading earlier in China outside of Wuhan, and later globally, providing immunity from the later appearing high-virulence strain (HVS). Following its spread into Wuhan, cumulative mutations gave rise to the emergence of a HVS, known as SARS-CoV-2, starting the COVID-19 pandemic. We model the co-infection by a LVS and a HVS, and show that it can explain the evolution of the COVID-19 pandemic and the non-trivial dependence on the exposure level to China and the population-size in each country. We find that the LVS began its spread a few months before the onset of the HVS, and that its spread doubling-time is $\sim 1.59 \pm 0.17$ times slower than the HVS. Although more slowly spreading, its earlier onset allowed the LVS to spread globally before the emergence of the HVS. In particular, in countries exposed earlier to the LVS and/or having smaller population-size, the LVS could achieve herd-immunity earlier, and quench the later-spread HVS at earlier stages. We find our two-parameter (the spread-rate and the initial onset time of the LVS) can accurately explain the current infection levels ($R^2 = 0.74$; correlation p-value (p) of 5.2×10^{-13}). Furthermore, countries exposed early should have already achieved herd-immunity. We predict that in those countries cumulative infection levels could rise by no more than 2-3 times the current level through local-outbreaks, even in the absence of any containment measures. We suggest several tests and predictions to further verify the double-strain co-infection model, and discuss the implications of identifying the LVS.

COVID-19 is thought to be a zoonotic disease, which currently spreads by human-to-human transmission[1, 2]. Human mobility mediates the geographic spread of the disease between countries through ground and air transportation. Given the origin of COVID-19 in China, it is therefore expected that traffic from China to other countries would drive the initial epidemic spread world-wide (e.g. [3]). In particular, it is expected that outbound mobility levels from China into other countries (i.e. the numbers of incoming passengers from China) should predict the risk/exposure level of the disease-spread into these regions. Here we make use of exposure-level, defined as the mobility level and risk assessment provided by the GLEAM epiRisk module (e.g. Refs. [4, 5, 6]), and compare it with the infection level (see Fig. 1a), measured either by the number of confirmed COVID-19 cases or confirmed COVID-19 deaths (and later accounting for normalization by testing levels and/or age-structure; see Methods).

We find that shortly after the beginning of COVID-19 spread, highly exposed countries show a linear positive correlation between the infection level (as measured by the number of COVID-19 confirmed cases), and the mobility-exposure level (given by the epiRisk module, based on flight-traffic data) at early times ($p = 5.5 \times 10^{-8}$; for countries with first cases by 80 days after December 1st 2019; see additional discussion of the early spread in Methods), but show no correlation with the population size ¹. However, at a later time (day 216), the same countries show a negative correlation between the infection-level and the mobility-exposure level ($p = 0.014$, $\rho = -0.66$; and $p = 0.006$; $\rho = -0.7$, for age-normalized deaths), as can be seen in Fig. 1a. Moreover, we find that at the later

*Corresponding author: hperets@physics.technion.ac.il

¹Note that unless otherwise stated, all correlations discussed are of the logarithms of the parameters

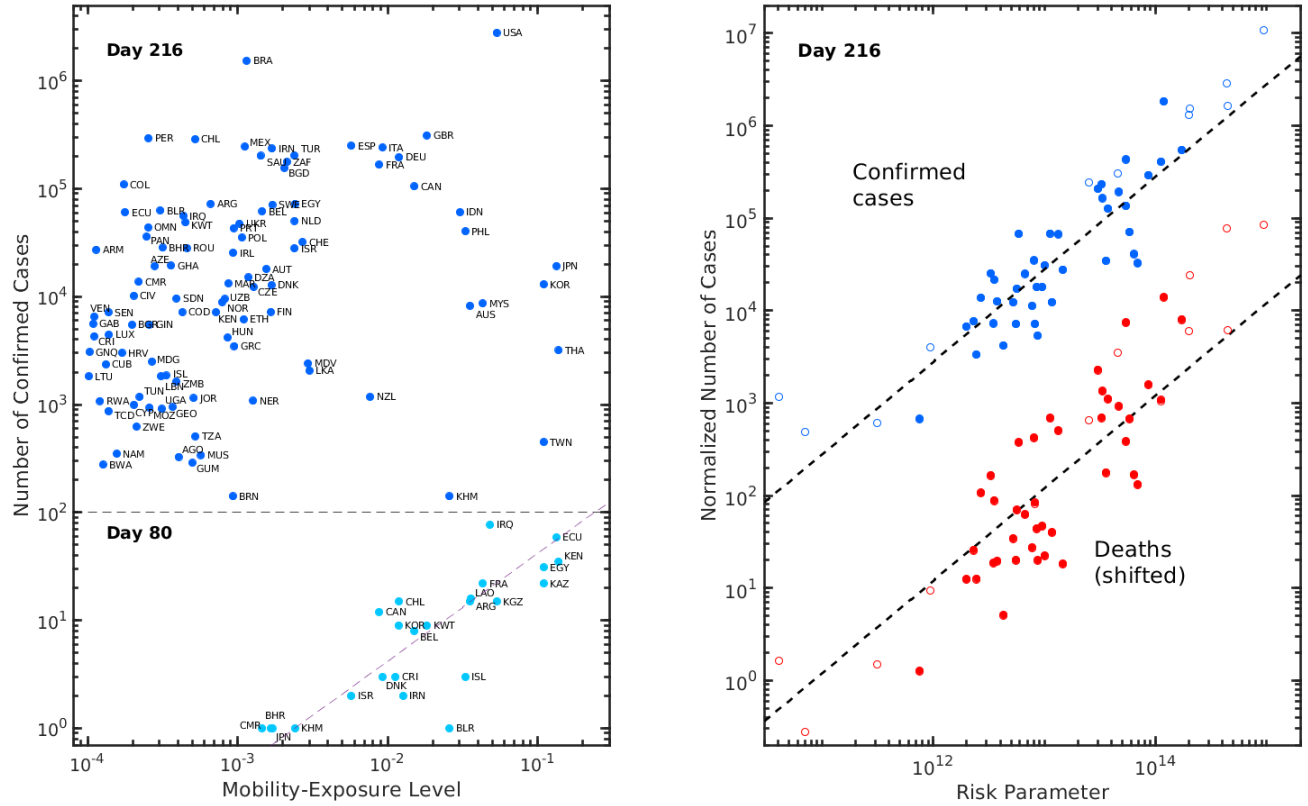


Figure 1: Left: Comparison of the number of confirmed cases as a function of the exposure level, at early (80 days since December 1st2019) and late (216 days) times. Early on highly exposed countries are infected earlier, and show an increasing number of cases as a function of the mobility-exposure level. At late times the same countries show a *negative* correlation between the infection-level and the mobility-exposure level. Right: The testing-normalized numbers of confirmed cases and testing (and age) normalized numbers of deaths (shifted down by a factor of ten for clarity), as a function of the risk-parameter (see definition in the main text) showing a direct linear relation as expected in the double-strain co-infection model. The least-populated and highest-populated countries (marked by open circles) were not used in deriving the models parameters as to assure they do not bias the results (see methods). Nevertheless, as can be seen, these too are also consistent with the model. See methods for a similar figure with annotated countries names (not shown here for clarity).

time point, the infection-level for all countries in our sample is positively correlated with the population-size (Pearson test; $p = 1.2 \times 10^{-9}$, $\rho = 0.55$ for confirmed cases; and $p = 2.8 \times 10^{-13}$, $\rho = 0.63$ for age-normalized deaths). While the early direct correlation of cases with exposure supports exposure-level risk estimates, our attention was drawn to the unexpected negative correlation between the infection-level and exposure-level at later times, and the correlation with population size at a relatively early epidemic stage where the numbers of cases are far below the levels necessary for herd immunity to take effect.

Here we show that the existence of an earlier low virulence (ancestor) strain (LVS) providing immunity from the HVS, could explain the current global distribution of the COVID-19. We propose that the LVS spread from a location in China, but far outside Wuhan, and later gave rise to the emergence of a the highly virulent strain (HVS), known as SARS-CoV-2. The exact location of the initial LVS's outbreak is unknown, but is likely to be in one of the least HVS-infected regions in China, and possibly close to Vietnam, given the very low infection levels observed there (see also section 1.3).

Previous studies explored epidemic models of co-infection and cross-immunity by multiple viral strains[7, 8, 9, 10, 11, 12, 13, 14, 15]. The emergence of highly pathogenic respiratory viruses from low-pathogenic ancestors has been previously described[16], and cross-immunity between respiratory low-virulence strains and high-virulence strains has been described as well[17, 18, 19]. In these cases the LVS effectively gave rise to a natural live vaccine, leading to a mild-to-asymptomatic presentation of the HVS disease later-on [17, 18, 19]. In the case of SARS-CoV-2, mutation rate was found to be slow, at the level of 1.3×10^{-3} substitutes per site per year[20], slower than the Influenza A/H1N1 pandemic virus which mutates at 4.4×10^{-3} substitutes per site per year[21]. Unlike the "antigenic shift" in fast mutating viruses that thwarts cross immunity and requires yearly vaccination[22], a slower mutation rate may allow for the co-existence of closer strains, retaining cross-immunity, or cross-response[23, 22]. We suggest that the mutation(s) that occurred in the LVS leading to the emergence of COVID-19 affected the pathogenicity and transmissibility of the virus, but not antigenicity. To date, genetic sequencing has shown that SARS-CoV-2 mutated and diverged into several genetic groups[24]. Mutations affecting pathogenicity have been previously described for SARS-CoV-2[25], but currently all identified groups, to the best of our knowledge, are highly virulent, inconsistent with the expected characteristics of the proposed LVS. Mutations affecting transmissibility, but not antigenicity, have been previously suggested for other respiratory pandemics[22]. Specifically, mutations effecting transmissibility and pathogenicity were previously described for the Ebola pandemic of 2013-2016[26, 27]. Our proposed model therefore follows similar pathways found earlier in other viruses.

As in those cases of cross-immunizing strains, our model suggests that although both strains share cross immunity, the earlier strain has mild to no clinical symptoms, leading to it being yet-unrecognized and not raising world-wide alarm. However, when this LVS reached Wuhan, an accumulation of mutations occurred that gave rise to the emergence of a faster spreading and more pathogenic HVS, namely the currently identified SARS-CoV-2, which then spread from Wuhan to China and the rest of the world. Our model suggests, given the delay between the onset of the LVS and the HVS, that countries with high exposure to China could achieve earlier LVS induced herd-immunity (or near herd immunity), leading to earlier quenching of the HVS spread. In countries that have lower exposure to China the LVS achieves herd immunity only later, or might not achieve herd-immunity at all before the HVS becomes widely spread. In these countries a larger COVID-19 pandemic would be expected. Such dynamics can explain the observed negative correlation between traffic from China and the infection-levels in each country. Similarly, since herd immunity is dependent on the percentage of population that is infected by a virus[28], less-populated countries could achieve herd immunity of the LVS earlier on, leading to the positive correlation between COVID-19 infection-level and the population-size.

As we show below, in our double-strain co-infection model, the overall infection level for any given country is prescribed by the mobility-exposure level of the country and its population-size, where the specific weight of these two parameters is determined by the spread rate of the LVS. Using a linear-regression analysis we can identify the LVS spread rate, and explain the current infection levels for the countries in our sample, which extends over four orders of magnitude in the number of cases. In Fig. 1b we show the correlation between the number of (testing-normalized) confirmed cases (see Methods) and the risk-parameter, which we define below (adjusted $R^2 = 0.73$; F-statistics vs. a constant model score of 112, with p-value of $p = 5.2 \times 10^{-13}$; 41 observation points, 39 degrees of freedom; using the 'fitlm' function in Matlab[29]) allowing us to explain (and predict) the past (and future) dynamics of the pandemic.

In order to study the dynamics of two-virus spread with cross antigenicity, we utilized the well-established Susceptible-Infected Recovered (SIR) model [30] (see ref. [28] for an overview). We extended the SIR model to the case of two circulating cross-immunizing viruses (i.e. a person infected by one of the viruses is immune to infection from the other virus, both during and after the infection; see Methods and related previous studies of co-infection and the spread of multiple strains[7, 8, 9, 10, 11, 12, 13, 14, 6]). In the model we track the number of people

susceptible to both viruses (S), the numbers of people infected with the LVS, or with the later-emerging HVS, (I_L and I_H , respectively), and the number of people who have had contracted either strain, and then recovered or died (R_L and R_H , for the LVS and HVS, respectively). It is assumed that the total population in a given country $N = S(t) + I_L(t) + I_H(t) + R_L(t) + R_H(t)$ is fixed, and that those who have recovered from either virus are immune to both. As we discuss later on, we assume the precursor LVS spreads slower than the HVS. The probability of a person contracting a given virus at time t is proportional to the fraction of susceptible people among the total population I_H/N for the HVS, and I_L/N for the LVS. These assumptions lead us to a set of five ordinary differential equations for $S(t)$, $I(t)$, and $R(t)$:

$$\frac{dS}{dt} = -\beta_L \frac{S(t)I_L(t)}{N} - \beta_H \frac{S(t)I_H(t)}{N} \quad (1)$$

$$\frac{dI_L}{dt} = \beta_L \frac{S(t)I_L(t)}{N} - \gamma_L I_L(t) \quad (2)$$

$$\frac{dI_H}{dt} = \beta_H \frac{S(t)I_H(t)}{N} - \gamma_H I_H(t) \quad (3)$$

$$\frac{dR_L}{dt} = \gamma_L I_L(t) \quad (4)$$

$$\frac{dR_H}{dt} = \gamma_H I_H(t) \quad (5)$$

Here, $\gamma_L, \gamma_H \geq 0$ are the recovery/death rates and $\beta_L, \beta_H \geq 0$, the transmissibility parameters, measure the likelihood of transmitting the LVS and HVS, respectively, when an infected and a susceptible person come into contact. Here we make a simplifying assumption that a person can transmit the virus to other people for a total period of 14 days, before the person recovers (or dies). While the LVS has low/negligible-virulence, as discussed earlier, and is assumed, for simplicity that infected people recover at the same rate from both viruses. In both cases we assume for simplicity that no measures to slow or stop the spread of either virus have been implemented. As can be seen in Fig. 1b, our results provide an excellent fit for the data, even without considering any containment measures, suggesting that the latter might have only a relatively small effect on the final infection level outcome, although they could give rise to an overall slower progression of the pandemic (see Methods for further discussion of our assumptions). Examples for the evolution of such double-strain co-infection model dynamics are shown in Fig. 2.

As can be seen in Fig. 2 (see also methods for further details), the larger the initial spread of the LVS (LVS seed-infection) at the time when the HVS began spreading in a given country, the faster the LVS achieves herd-immunity, thereby also quenching the fast spread of the HVS, and giving rise to a lower HVS-infection level. Since herd-immunity requires a population-wise infection-level (e.g. $\sim 1 - 1/\beta\gamma$ of the population; see [31, and references therein]), more populated countries take longer to achieve herd-immunity for the same seed-infection (Fig. 2c & Fig. 2d). This allows for more time for the spread of the HVS, giving rise to an overall higher HVS infection-level. Since the spread of the HVS is faster than that of the LVS, the former (HVS) can even outrun the latter (LVS), in which case the HVS will not be limited by the LVS, and will only be limited by the population size (See Fig. 2c).

Based on the double-strain co-infection model we can now explain the observed distribution of COVID-19. The LVS outbreak began in China, far from Wuhan. Shortly after the LVS reached Wuhan, and before the LVS achieved herd immunity in this region (e.g. similar to the case in 2b), cumulative mutations lead to the emergence of the HVS in Wuhan, initiating the initial exponential spread of the HVS strain and its outbreak in China and later worldwide. At later time the HVS spread slowed down and almost completely stopped once the LVS achieved herd-immunity, quenching the exponential spread of both the LVS and the HVS. Note that in the model the total *cumulative* number of infections can still slowly rise by additional factor of 2-3 after herd-immunity is achieved (which approximately corresponds to the time when the peak in active-cases is seen; see Fig. 2). Throughout China the LVS had more time to spread before the spread of the HVS, leading to immunization of an even larger fraction of the population (compared to Wuhan and Hubei region) before the spread of the HVS from Wuhan, leading to an even lower COVID-19 infection rate in China outside Hubei province.

The global spread of the LVS followed the exposure level, with more exposed countries receiving LVS-infected people earlier (strong negative correlation between the logarithm of the exposure level and the day of first infection (see methods and Fig. 4; $p = 6.5 \times 10^{-10}$). If the LVS and HVS spread rates were the same, the spread of both strains would have followed the exact same dynamics, and the delay between the arrival of the LVS and the HVS would be constant. Consequently, a fixed ratio of LVS-infection level to HVS-infection level in different countries (with similar population sizes - see below) would be expected, leading to an overall similar HVS-infection levels.

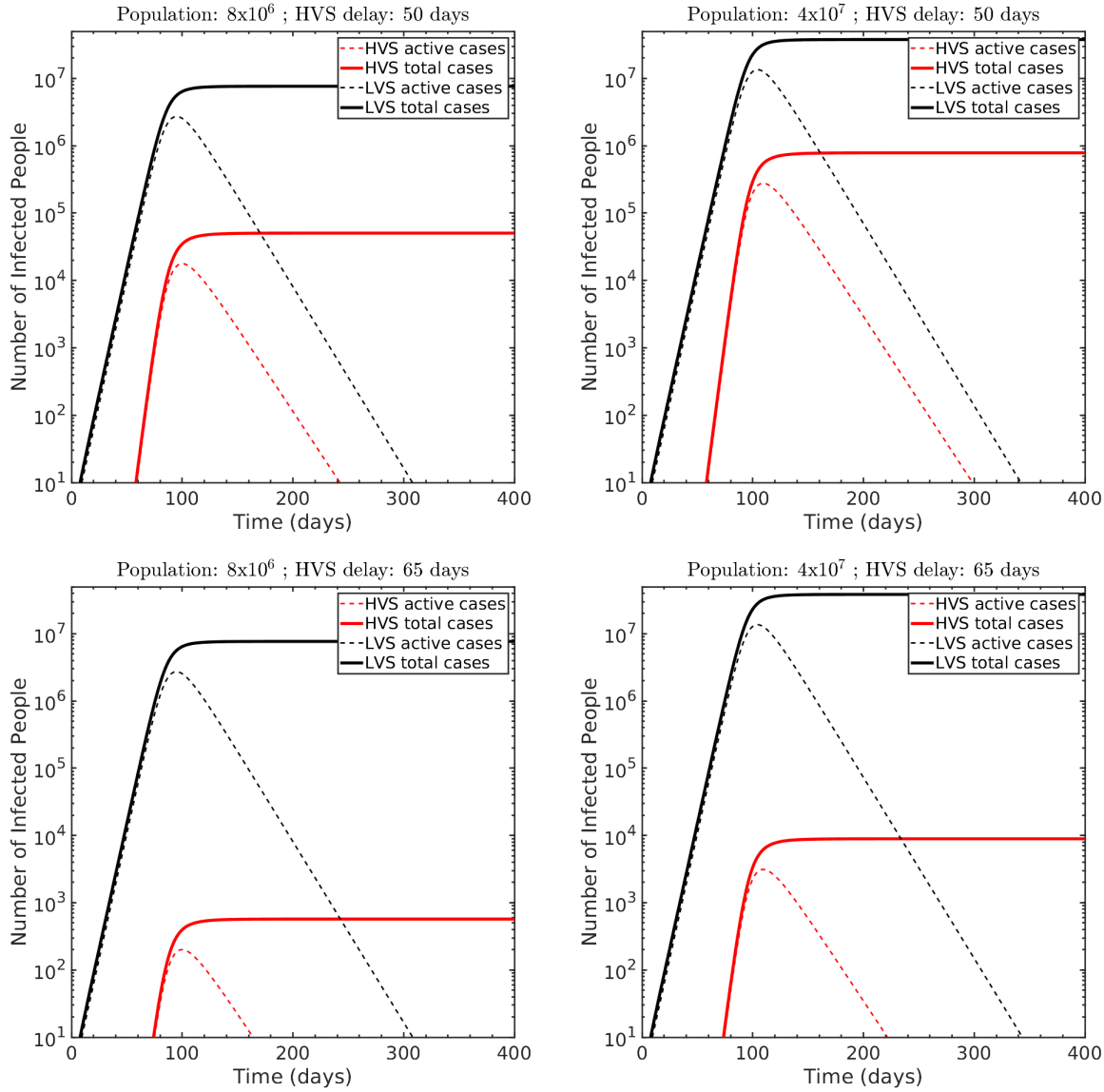


Figure 2: The results of a simple SIR co-infection model, showing the evolution of two viruses which share cross-immunity, where the slower-spreading one (the LVS, in our case), begins spreading earlier. We show four different cases as to exemplify the dynamics. Top (bottom) left : The LVS begins spreading 50 (65) days earlier than the HVS in a population of 8×10^6 people. Top (bottom) right: The LVS begins spreading 50 (65) days earlier than the HVS in a population of 4×10^7 people. As can be seen, the longer the delay between the onset of the LVS and the HVS (which is a function of the mobility-exposure level, see Eq. 8), and the smaller the population, the spread of the HVS is quenched earlier, leading to an overall lower infection-level of the HVS (see test for further discussion).

However, the infection levels in countries of similar population sizes differs between lower and higher exposures (e.g. compare France and Thailand, Taiwan and the Netherlands in Fig. 1). Therefore, the observed differential infection level, requires the LVS transmissibility level to be lower (consistent with our finding below). In this case, as shown in Fig. 2 and discussed in the Methods, the HVS spread can outrun the LVS spread and decrease the relative LVS to HVS infection levels with time. The LVS spread to less exposed countries started later on, allowing for a shorter time frame for the faster HVS spread to outrun the LVS spread, giving rise to a differentiated increase in the ratio of HVS-to-LVS infection with decreasing exposure-level. For a given population size this ratio determines the final HVS infection level (compare Figs. 2a and 2c).

Another important parameter that determines viral spread in the double-strains co-infection model, is population size. As discussed above, the time to achieve herd immunity depends logarithmically on the population size. Therefore, countries with smaller population size would approach herd immunity faster, allowing less time for the HVS spread to outrun the LVS-spread before its exponential growth is quenched.

Considering the exposure-level - dependent HVS to LVS infection ratio, the population size, and the initial delay between the onset of the LVS and the onset of the HVS in China, we expect the HVS infection level to have a specific dependence on these parameters (as we derive in details the Methods), given by

$$\log_2(I_H^{\max}) = [(t_0^L - t_0^H)/T_H] + \left(\frac{T_L}{T_H}\right) \log_2 N - \left(\frac{T_L}{T_H} - 1\right) \log_2 c_{exp}, \quad (6)$$

where $I_H^{\max}$ is the expected maximal infection level of the HVS, t_0^L and t_0^H are the onset times of the LVS and the HVS, respectively; and T_L and T_H are the doubling times for the LVS and HVS, respectively. In particular, we expect a specific relation between the dependence on the population size, and on the doubling times ratio (T_L/T_H compared with $T_L/T_H - 1$; see Methods). Linear regression can therefore allow us to identify the specific dependence. Using the data on confirmed cases we find best fit linear regression models (for Eq. 6) of $T_L/T_H = 1.31 \pm 0.23$ (Adjusted R^2 of 0.42). Although already providing highly significant results ($p = 5.3 \times 10^{-9}$; Pearson correlation test), such data are noisy due to differences in the testing level. When analyzing the testing-normalized confirmed cases (see methods) we find best fit linear regression models (for Eq. 6) of $T_L/T_H = 1.59 \pm 0.17$ (Adjusted R^2 of 0.74). Using the age-normalized deaths data we find $T_L/T_H = 1.57 \pm 0.32$ (Adjusted R^2 of 0.41), or, when also age *and* testing-normalized data, giving $T_L/T_H = 1.72 \pm 0.26$ (Adjusted R^2 of 0.62). All of these are consistent with each other and the model. Hereafter, we adapt the most significant result, $T_L/T_H = 1.59$, as our fiducial ratio. Correlation tests between the risk-parameter, χ , which we define (following Eq. 6) as

$$\chi = \left(\frac{N}{2^{(t_0^H - t_0^L)/T_L} c_{exp}^{1 - T_H/T_L}} \right)^{T_L/T_H},$$

and the number of testing-normalized cases (age and testing-normalized deaths) for $T_L/T_H = 1.59$ give $p = 5.2 \times 10^{-13}$, 3.8×10^{-10} and $p \ll 1e - 20$ ($p = 1.8 \times 10^{-10}$, 1×10^{-7} and $p = 1.4 \times 10^{-7}$), for the Pearson, Kendall, and Spearman tests, respectively.

Taking the mobility exposure-level in China to be unity, by definition, and plugging the population size and the infection level, we can then find the delay time between the onset of the LVS and the HVS. If we assume a doubling time of 2.7 days (in the absence of any containment measures), consistent with observed spread rate at the earliest days, we find $(t_0^L - t_0^H) \sim 86$ days, suggesting the initial human infection by the LVS began in China around September. This, however does not account for non-diagnosed infections, population structure and inter-China mobility that may slow transmission, and it is possible that the delay might somewhat differ. If we assume only ~ 0.1 of of the infections are diagnosed we get ~ 77 days.

In Figs. 1c and 1d we show $I_H^{\max}$ for both the testing-normalized confirmed cases and for the age and testing-normalized deaths, as a function of the risk-parameter defined above. In fact, low and high population countries, not used to derive the model parameters as not to introduce potential bias (see Methods) also well fit (including them in the statistical score gives $p = 7 \times 10^{-22}$ (Pearson) for the same model parameters), and the model shows an excellent agreement with observations, over five orders of magnitudes in infection level. We find that that in most countries in our sample the spread of the LVS should achieve herd-immunity level, and quench further exponential spread of the HVS, much before the HVS achieves herd-immunity infection-level. We find low dispersion around the expected values from the model (up to a factor of a few for the testing-corrected confirmed cases), leaving relatively little room for other parameters beside exposure level and population size in affecting the final outcome, as we further discuss below. A wider dispersion can be seen in the age-corrected deaths. However, the amplitude of the dispersion grows with the number of deaths/confirmed-cases, suggesting a relation to the load on the health system affecting treatment, and thereby the number of deaths. A further inquiry on this issue, though important, is beyond the scope of this paper and will be explored elsewhere.

Our model and analysis of the current statistical data supports the existence of a preceding LVS. They explain and predict upper limit for infection-levels as a function of the mobility-exposure to China and the population-size. In particular, we find that more exposed countries the spread of the LVS is already at or close-to herd-immunity level, which explains the puzzling observed low-level infections in these countries, and the direct, but non-linear dependence of infection level on population size, which is otherwise unexpected given the low infection level in respect the those required for herd-immunity. However, additional independent tests could further support or refute the model. These can be divided between biological tests and demographic one, as we discuss in the following.

Cross antigenicity between the LVS and COVID-19 is an essential part of our model and can be either humoral or cellular. If cross antigenicity relies on cross reactive antibodies, it is possible that antibodies to LVS will be detected by serological testing for SARS-CoV-2, already employed to some extent in several countries[32]. In particular countries which appear to have achieved LVS herd-immunity should have shown a large fraction of the population to be sero-positive, in contrast with the currently directly measured lower per-population HVS infection-levels[32, 33]. However, current serologic testing is optimized for HVS, and might be less efficient for detection of the LVS. Recent findings at the levels of a few percents up to 30 percents sero-positive, i.e. tens to hundred times larger than the infection rates inferred from the confirmed cases, but still considerably lower than required for herd-immunity. These likely reflect low-testing levels in most countries, suggesting that actual HVS-infection levels are typically much higher than inferred from the reported cases, consistent with our use of testing-normalization in our analysis (see Methods). It also suggests that none of the currently identified genetic groups of SARS-CoV-2 could be the LVS proposed here. Moreover, it is important to note, that it is currently unclear whether the antibodies detected in the currently available serologic tests are indeed protective antibodies, and therefore even in the cases of humoral cross antigenicity, LVS strains might not be detected at all by SARS-CoV-2 serologic tests. We therefore suggest to make use of a more direct and accurate method to test the existence of LVS antibodies, by employing a viral micro-neutralization testing of the HVS on serum from a sample of people (preferably from highly exposed countries, for which the majority of the population should already have been infected by the LVS) who are found to be negative for the SARS-CoV-2 HVS in serologic tests. These should be able to identify antibodies reaction to the HVS otherwise undetected by currently employed serologic test, in a similar manner as used to test acquired immunity to a high pathogenic flu virus following vaccination for a less pathogenic flu virus[34]. To the best of our knowledge only one micro-neutralization study (with limited statistics) has been done on SARS-CoV-2 to date[35], not identifying antibodies in the serum of people who were found to be negative for the SARS-CoV-2 HVS in serologic tests.

A likely possibility is therefore that cross antigenicity between the two strains might be dependent on *cellular* immunity, rather than humoral one. For example, previous studies on avian influenza HVS H5N1 protection by the LVSs H9N2[17, 18], H1N1, or H1N2[19] showed that the immunity was cellular rather than humoral immunity, and H5N1 antibodies were not found in chickens that were previously infected by low virulence strains[17, 18, 19]. In fact, very recently, and after the completion of our analysis Sekine et al. [36] detected SARS-CoV-2-specific T cells were detectable in antibody-seronegative family members and individuals with a history of asymptomatic or mild COVID-19. This could be consistent with our suggestion of cellular cross-immunity and the possible widespread immunity due to the LVS. Our model suggests that mapping T-cell response test among a current sample of people with no known infection in most countries and especially in highly exposed countries (e.g. China, Vietnam, Thailand) should show the majority to already be immune, inconsistent with the known low-level infection level in these countries in respect to the population size, verifying our results. Although, at this stage, one can not exclude acquired immunity through past infections from other viruses, however, such immunity level should not follow the specific correlations expected from our model and explaining the current global distribution.

Assuming a LVS SARS-CoV-2 is identified and a specific serologic or mapping T-cell response test is developed for it, tests of blood samples taken across China (or in other countries achieving LVS herd-immunity) around October-November, i.e. before the December 2019 HVS outbreak, but at the point where the LVS should have already been widely spread, could identify positive cases, in particular outside Hubei province. In fact, the study of samples taken at different times would potentially allow for identifying the LVS spread, where positive fractions should gradually increase from September until November. Similarly, early evidence for SARS-CoV-2-like viruses found before the known COVID-19 outbreak in December 2019 could further support our finding, e.g. through finding evidence for SARS-CoV-2 in the sewage, much before the COVID-19 outbreak began at the respective country[37], could provide similar clues of the early spread of the LVS. In particular, a much earlier spread of the HVS should have been easily identified due to the large number of patients expected, inconsistent with the data, while a wide spread LVS, could be consistent with such findings.

Biological tests could provide optimal smoking-gun signatures of the LVS spread, but are depend on the unknown genetic similarity of the LVS and the HVS, the type of serologic or mapping T-cell response tests and the type of

cross antigenicity. A different, sequence-independent statistical test approach can be used to test for LVS spread. People visiting China in the 1-2 months before the HVS outbreak (but not afterwards), might have been infected by the LVS during their visit, thereby acquiring immunity to the HVS. HVS infection-level among such people should therefore be lower than that of the background population in their home country, in particular in COVID-19 highly infected regions. Therefore testing COVID-19 prevalence among people visiting China in September to November 2019 could confirm the presence of a protective LVS strain independent of its genomic sequence, but requires non-trivial data collection due to the large travel and health information needed for this statistical analysis.

The existence of a LVS providing cross-immunity to the COVID-19 HVS strains has far-reaching implications. First, it proves the existence of acquired viral immunity and the potential for developing an efficient vaccine, given the immunity provided by the LVS as an effective live-vaccine. Biological identification of the LVS could be used as a highly advantageous starting-point for developing a safe COVID-19 vaccine, although modifications preventing the occurrence of new pathogenic properties of the natural virus will be essential. Moreover, although the LVS might protect from further infections in already “protected” countries, it is yet unknown how long such immunity protection lasts (see e.g. discussion of limited immunity in Ref. [38]). Lastly, the HVS SARS-CoV-2 might at some point change its structure to generate a novel HVS that would not be cross immune with the previous strains.

Biological identification of the LVS would also enable the development of specific serologic tests and/or mapping T-cell response tests, that can then be used to measure and directly validate the existence of herd-immunity in a given country, and thereby direct the application of containment measures and opening strategy plans.

Non-homogeneous geographical transmission and/or non-homogeneous transmission among less/disconnected communities can give rise to pockets of less immune populations within largely-immune countries, explaining the possibility of local outbreaks of the HVS. More generally, our simplified model considers countries as independent units, while intra-country dynamics could lead to a down-scaled version of the model. Even in countries in which the early LVS infection provided partial protection, the protection might not be homogeneous. Regions and/or communities more exposed to international traffic would be infected earlier, leading to initially higher infection levels, but eventually show lower infection-level at later times, since the same regions/communities also acquired higher level of immunity through larger prior LVS infection. More isolated regions and/or remote/disconnected communities might eventually be at higher risk and could have higher infection levels compared with the overall country population. Generally, regions/communities having earlier infections are likely to experience overall lesser cumulative levels of infections, compared to same-size communities infected later, and larger isolated communities are generally likely to experience higher infection level. Similar intra-country analysis (e.g. comparing the exposure levels and population-sizes in different states in the US, would suggest some of the more populated states in the US are still likely to experience higher infection levels), could serve to predict the intra-country dynamics of the COVID-19 pandemic. Furthermore, COVID-19 can still affect immune deficient people everywhere, as these are not protected by the LVS, although herd-immunity significantly lowers their chance of infection.

It is possible, and even likely that in a fraction of the cases the LVS only provides partial immunity, allowing for HVS infection, but leading to a less virulent form of COVID-19. In such a case we might expect a higher fraction of newly identified COVID-19 patients to be, on average, more symptomatic during the early phases of the pandemics when it still spreads exponentially, before the LVS achieves her-immunity. At these early times the LVS has not yet infected the majority of the population and newly HVS-infected people are not likely to have been previously infected by the LVS, and be partially immunized. After the LVS infected a large fraction of the population, new HVS-infections are far more likely to be of previously LVS-infected people, who already acquired partial immunity. We might therefore expect a lower fraction of asymptomatic cases, and a higher morbidity rate during the early exponential growth of the HVS, in comparison with later times after sub-exponential-growth and later decay in the number of cases is observed. We do note that the improvement in COVID-19 treatment in time due to the learning curve of the health systems might give rise to lower morbidity at later times too, but it is less likely to affect the asymptomatic to symptomatic ratio, nor should it be related to the transition from exponential to sub-exponential growth. Note, that in a case of only partial immunity, the infection-level, in terms of cases, could somewhat increase beyond the level described above, the transition to sub-exponential growth, but the infection-level in terms of deaths is likely to be far less affected.

Our results indicate that the infection level depends on the exposure level and the population size. In particular there is relatively little dispersion in the observed infection levels in respect to the basic model fit. This suggests that although the widely varied containment measures applied by different countries[39] might be important in slowing the virus spread to some extent, and possibly allowing the health system to better accommodate more serious cases, the specific and different measures taken by each country eventually had little effect on the number of infected people in a given country. In particular, it is difficult to understand how would the different measures taken at different times could give rise to the type of correlations with exposure-level and population sizes which we identify.

Our results would similarly suggest that different opening strategies after lock-downs and social distancing would not considerably affect the overall infection levels, which is effectively dependent on the double-strain dynamics and not the containment measures responses (or the lack of them). While it is not clear whether any of the currently used containment measure had a significant effect on the overall number of infected people, it is important to note that in principle, measures which could be more effective for eradicating the HVS compared with the LVS should be prioritized. Such measures may change the relative LVS to HVS infection levels, and would provide the same effect as introducing a higher per-population exposure, and in turn constrain the maximal HVS infection level to lower values. Current contact tracing, for example, is focused on HVS infections, while not affecting the LVS spread; similarly, quarantine/isolation of HVS-infected people decreases the HVS-transmissibility. Social distancing, however, affects both the LVS and the HVS, and might even be harmful under some conditions, as it lowers the transmissibility parameter β for both the LVS and the HVS. Given that the doubling time and β are related through $T = \log(2)/\log(1 + \beta)$, one gets $T_L/T_H = \log(1 + \beta_H)/\log(1 + \beta_L)$. Decreasing both β_H and β_L by the same factor increases T_L/T_H allowing the HVS to outrun the LVS faster, i.e. leading to an overall higher HVS infection-levels. However, this is only true as long as the reproduction number is bigger than unity and growth is exponential, if the containment measures are sufficient to drive the reproduction number below one, the exponential spread of the disease is stopped in any case, and than these measures are indeed helpful.

Fast mutating viruses might mutate too fast as to give rise to long-lasting cross-immune strains. The slow mutation rate of SARS-CoV-2 could be the reason allowing for LVS and HVS co-infection. Nevertheless, any new low-virulence viruses are likely to go through several mutations before becoming virulent. This would suggest that continuous monitoring of viruses mutations in the population could be used both to identify the potential progress into virulent forms, as well as serve for development of new vaccine, once a virulent virus forms. In particular, sample collections of non-virulent viruses and their sequencing done continuously could serve as key input in the onset of a new epidemic, allowing for immediate back-tracking of the new-virus development, and the possible use of earlier non-virulent strains as an advanced starting point for development of a vaccine.

Finally, although developed in the context of the COVID-19 pandemic, similar multi-strain analysis and the statistical identification of co-infection described here can be used for the study of other pandemics.

1 Methods

1.1 Data sources

We have used several data sources for this analysis, described herein .

COVID-19 cases and tests and population sizes: The numbers of confirmed COVID-19 cases, deaths, and number of tests, as a function of time, as well as the data for population sizes were taken from the our world in data (OWID) website (www.ourworldindata.org), which (COVID-19) data is based on the EU-CDC data (<https://www.ecdc.europa.eu/en>), which we downloaded on 4/7/2020. For verification, we also compared the data from this site to that in the worldometer website (<https://www.worldometers.info/coronavirus/>) and find it consistent with the data we used, besides the number of tests for Peru, which is much higher in the worldometer website. For consistency, in our analysis shown here we only make use of the OWID data, which also provide the full data as a function of time since the beginning of the pandemic, and not only the last few days as done in the worldometer data. Nevertheless, we did make the same analysis on the worldometer data for verification (the data also contain testing data on several countries not available in the OWID data), and found fully consistent results, i.e. the same doubling time ratio, and at a similar level of significance, further verifying the data reliability and consistency.

Mobility-exposure level: The main analysis is based on exposure data derived by the open tool web-platform EpiRisk module at <https://epirisk.net/> taken on 3/6/2020. As described on the web-site, EpiRisk is a computational platform designed to allow a quick estimate of the probability of exporting infected individuals from sites affected by a disease outbreak to other areas in the world through the airline transportation network. EpiRisk is part of the the Global Epidemic and Mobility project (GLEAM; <https://www.gleamproject.org/>) which combines real-world data on populations and human mobility as to inform epidemic models and was already used for studying COVID-19 spread[6]. These data provided us with exposure level estimates for over 100 countries with the highest air-transport mobility rates incoming from China. EpiRisk mobility data also accounts for seasonal differences, and we considered both risk model in the Autumn (August-October) and Winter (November-January), generally providing very similar results; the results shown in the paper are for the Winter mobility results, at which time most of the infections were likely to occur.

In order to enable an easy reproduction of our results we provide the EpiRisk mobility-exposure data (for the Winter season). We also provide the risk-parameter we derive for each country in the same EpiRisk mobility-

exposure file. In addition we provide the data on the number of COVID-19 cases, deaths, population-size, number of tests and age-structure from the “Our world in data” website, taken on July 4th 2020.

1.2 Data filtering

In our analysis we considered data only for countries for which our above mentioned data sources provided all relevant information, i.e. number of confirmed cases, number of deaths, number of tests and mobility-exposure levels, and population sizes (the latter was available for all countries). We then made use of several filtering and normalization methods to minimize data noise level, as described below.

1.2.1 Age and testing normalization

The exact numbers of COVID-19 infections are unknown, since screening of the entire population has not been done in any country. We therefore use the numbers of confirmed cases and the numbers of COVID-19 related deaths as proxies for infection levels. However, each of these numbers is influenced by various factors, which are difficult to assess, and their exact relation to the actual infection level could differ to some extent in different countries, giving rise to effective noisy data. We made of two corrections in order to filter some of the main components.

COVID-19 related deaths are typically of patients which were earlier critical condition. These are more likely to have been identified and tested than asymptomatic patients, which might have never been tested. Confirmed deaths are therefore potentially more reliable proxy for the infection level. On the other hand the majority of the deaths are of people older than 70, and therefore differences in age structure in different countries will give rise to different ratio between the infection level and the number of deaths. We therefore age-calibrated the death rates in each country by dividing the number of deaths by the fraction of the population older than 70 years. Other potentially important affecting factor is the quality of the health system, where better health systems might be able to provide better treatment and decrease the morbidity level. Finally, even in high quality health systems, an overload of a large number of patients in critical conditions might affect the quality of treatment, potentially leading to higher morbidity rates in various countries. The two latter factors are more difficult to assess quantitatively, and are not accounted for in our analysis. We only considered countries for which our data source provided the age-structure.

The number of confirmed cases is likely to be less affected by the overall quality of the health system, but is more dependent on the number of COVID-19 tests done in respect to the population size, and the testing strategy. We therefore normalized the numbers of confirmed cases in each country by dividing them by the corresponding numbers of tests per population in these countries. Since our model predicts a relation between infection level and the population size, one might be concerned that inclusion of such a parameter which is related to population size could introduce an artificial correlation. We find that the tests per population show no correlation with the population size ($p = 0.15$; see also Fig. 3) when excluding the most ($> 10^8$ people) and least ($< 2 \times 10^6$ people) populated countries in our sample. In Fig. 5 we show the data for all countries in our sample, including these high/low population ones. As can be seen, these data for these countries is highly consistent with the model. In fact, when adding these countries to the analysis (but not the systematic outliers (see below) also shown in the figure) the results even further improve, and we find adjusted- $R^2 = 0.84$ and Pearson correlation test p-values of 4×10^{-22} , for the same model parameters derived without considering these countries.

1.2.2 Mobility-exposure level

Throughout our analysis we assume that each country is exposed to the viruses directly through traffic from China. While this is likely to be true for countries highly exposed to China, less exposed countries might be infected via secondary spread from other countries, even before being infected through direct incoming traffic from China. A detailed pandemic model which accounts for the traffic between each country and the epidemic evolution in time could potentially better account for such secondary exposed countries. This is beyond the scope of the current study, and we therefore only account for countries which were likely to be directly exposed to China. In Fig. 4 we plot the day of first confirmed infection for each country as a function of the exposure level to China. From Eq. 8 (see below) we expect a correlation between the logarithm of the exposure level and the initial infection in a given country. An overall strong correlation can indeed be found for the full data sample ($p < 10^{-17}$), but above ~ 100 days the correlation saturates, possibly suggesting that countries exposed to COVID-19 more than ~ 100 days after December 1st 2020, are more likely to have been exposed indirectly. Therefore, in our analysis we only make use of data from countries in which the first confirmed case has been identified at most 100 days after December 1st 2019. As can be seen a large dispersion can already be seen before 100 days suggesting that the exact identification of countries with experience direct or secondary exposure is difficult to discern without a

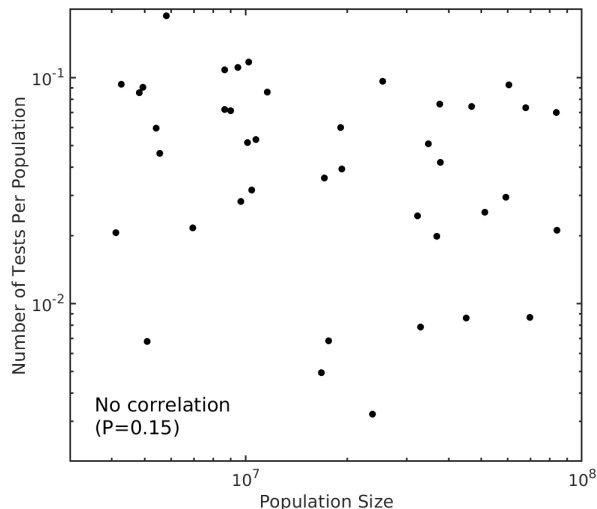


Figure 3: The relation between population size and the number of COVID-19 tests per population. As can be seen we find no correlation ($P = 0.1$) between these parameters.

more detailed model. Nevertheless, we find that even making use of the full data, i.e. even beyond 100 days, does not affect our conclusions, i.e. we can still identify (with significant statistical score, albeit less significant than those discussed above) model parameters consistent with those we find using the filtered data. In principle the initial infection day might be used (after appropriate analysis) as a phenomenological proxy for the exposure level, irrespective of the traffic data. This is beyond the scope of the current analysis and will be explored in later studies.

1.3 Systematics and outliers

Throughout our analysis we have made use of all data available to us, as described in section 1.1, and the statistical significance levels are based on analysis of these data. Nevertheless, as mentioned in the main text, some clear systematic effects and outliers can be identified in respect to specific countries. It is therefore important to discuss these systematics and the possible origin of outliers.

Our analysis makes use of exposure-level assessments based on air-traffic mobility from China. This may give rise to two systematic types of incorrect estimates of the exposure level. Underestimates for China neighboring countries, and overestimates of flight-transportation hubs.

China has long ground borders with neighboring countries allowing for considerable ground traffic mobility, which data are not available to us. Passenger trains operate from China to Vietnam, Mongolia, Kazakhstan, Russia and North Korea. Besides trains, border crossing points allow for ground mobility to the other neighboring countries, including Bhutan, Kyrgyzstan, Kazakhstan, Laos, Nepal, Tajikistan, Myanmar, India, Pakistan and Afghanistan. The exposure level of these countries are therefore likely to be underestimated. Increasing their exposure level could drive them to the regime of higher per-population exposure level, and constrain and lower their HVS infection level due to the LVS herd-immunity limit. We therefore did not include these countries in our main analysis.

Air-traffic data might not well account for flight-hubs, in which significant fraction of incoming passengers are commuting through on their way to other countries. In such cases the exposure level might be overestimated. In more populated countries, the addition of the commuting might not considerably affect the overall incoming traffic level. The only cases in which the annual traffic through the airport is at least approximately ten times larger than the population in the host country itself are Dubai airport in the United Arab Emirates, Singapore airport and Doha airport in Qatar (all three by a large margin from other major hubs; according to the Airport Council International world airport traffic and rankings). The exposure level of these flight-hubs are therefore likely to be overestimated. We therefore did not include these countries in our analysis.

Although not included in the main analysis, given the uncertainties, these countries do provide some indirect information. In Fig. 5 we show the same data as in Fig. 1, but now highlighting the expected “outlier” underestimated exposure China-neighbors and overestimated exposure flight-hubs countries. As can be seen the locations of these countries and the outlying position are consistent with the appropriate expectations, giving further (though indirect) support for the double-virus co-infection model. In particular, in a single HVS model, the neighbouring

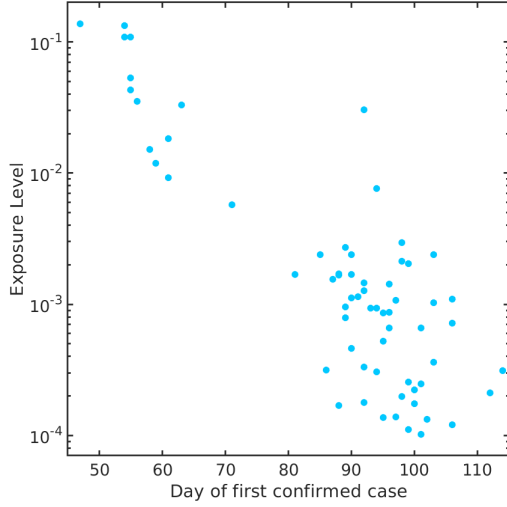


Figure 4: The day of first confirmed infection for each country as a function of the exposure level to China.

countries should be more exposed and show *larger* infection-levels and the flight-hubs should be less-exposed showing *lower* infection level, i.e. the opposite expectation in respect to the double-strain model, and inconsistent with the COVID-19 data.

1.4 Preceding low virulence strain and the co-infection SIR model

In order to demonstrate the basic aspects of the model and its observable expectations, we make use of the Susceptible-Infected-Recovered (SIR) model[30], often used to study the spread of infectious disease. We extend it to the case of two circulating cross-immunizing viruses (i.e. a person infected by one of the viruses is immune to infection from the other virus, both during and after the infection), following past studies of co-infection and the spread of multiple strains[7, 8, 9, 10, 11, 12, 13, 6]), as described by Eqs. 1-5 in the main text.

Let us now consider the dynamics of the epidemic in two possible scenarios (1) the HVS spreads in a given country earlier than the LVS, or the LVS spreads earlier, but the HVS spreads in a given country shortly after; and (2) The LVS spreads much earlier in a given country well before the spread of the HVS. In the first scenario the HVS spreads into the country before or shortly after the first arrival of the LVS. In this case (see 2b), although some fraction of the population becomes LVS-infected, the spread of the HVS outruns the LVS spread, even if the LVS began before (but not too long before, as that would lead to the second scenario) and HVS effectively spreads freely with little effect of the early LVS-spread. In particular, in such cases the HVS-infection level could infect the majority of the population, if no measures are taken, and until then its spread is independent of the country population size, and just grows exponentially with time, as if only a single strain exists. Note that the same evolution also describes the early stages in the second scenario, before the LVS reaches herd-immunity and affects the spread of the HVS.

In these cases, the infection-level at a given time should just be approximately

$$I_H(t) = 2^{(t-t_{exp})/T_H}, \quad (7)$$

where t_{exp} is the time of the initial infection (exposure) in the given country, which is a function of the exposure level to China, as we discuss below. The number of infected people arriving to another destination country from China is proportional to the number of infected people in China, and the fraction of those people who travel from China to that destination. The number of infected people in China at a given time (assuming no containment measures are taken) is $2^{(t-t_0)/T}$, where t_0 is the time of the initial infection in China, and T is the doubling time of a given virus. The fraction of travelers from China to a given country is proportional to the mobility exposure level, $c_{exp} = \alpha\mu$ to that country, where μ is taken from the GLEAM project (see section 1.1), and α is some constant unit calibration to account for the GLEAM data units. The time for the first infection (exposure) in a given country (t_{exp}) therefore follows the following relation

$$1 = 2^{(t_{exp}-t_0)/T} c_{exp},$$

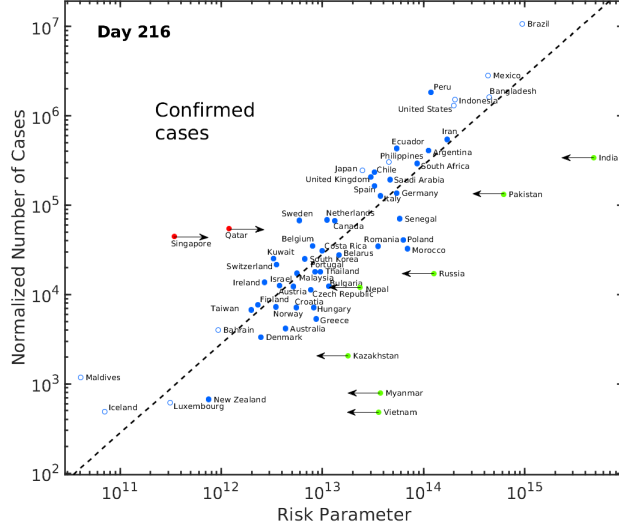


Figure 5: The testing-normalized number of cases as a function of the risk-parameter, including the expected outliers. China neighboring countries (shown in green) could potentially receive considerable ground traffic from China besides the flight-traffic, while only the latter is included in the exposure parameter. The risk-parameters of these countries are likely to be underestimated. Flight hubs (relative to population size; shown in red) are likely to have considerable incoming traffic from China of people who only transit through these countries. The risk-parameter of these countries is likely to be overestimated. The arrows emphasize that that the risk-parameters for hub-countries only serve as lower-limits and that the risk-parameters of neighboring-countries only serve as upper limits.

which can be written as

$$t_{exp} - t_0 = -T \ln_2 c_{exp}. \quad (8)$$

Plugging this result in Eq. 7 we get

$$I_H(t) = 2^{(t-t_{exp})/T_H} = 2^{(t-t_0)/T} c_{exp}, \quad (9)$$

In other words, at a given time we expect to see a direct linear correlation between the exposure level, and the number of confirmed cases at given country, at least until the majority of the population became infected and a herd immunity was achieved (by either virus). This is consistent with the results in Fig. 1. At early time the evolution follows the regular dynamics of a pandemic by a single virus, and allows us to find the calibration parameter α by simple linear regression on the data. At later times we see a different behavior, which can be understood by considering the second scenario.

In the second scenario, the LVS begins spreading in the community and the number of LVS-infected people grows exponentially following the same evolution as described for the HVS in Eq. 7. If the LVS spreads far earlier than the HVS it could have infected the majority of the population as to induce herd-immunity, and by the time of first HVS infection, the HVS can not spread. If the HVS begins its spread long after the LVS, but before the LVS achieved herd-immunity, both viruses initially grow almost unimpeded, until the LVS achieves herd-immunity, at which point both viruses spreads stop the exponential growth, slow down and achieve their maximum (See Fig. 2).

Longer delay between the initial spread of the LVS and the initial spread of the HVS therefore translates into lower number of HVS infected people and further constrains its spread up to a lower fraction of the overall population (compare Figs. 2b and d to Figs. 2a and c, respectively). For the same delay, countries with smaller population sizes will achieve herd-immunity sooner, allowing for a lesser number of HVS infections (compare Figs. 21 and b to Figs. 2c and d, respectively).

The introduction of a precursor LVS can therefore explain the otherwise puzzling findings of the highest exposed countries showing low infection levels at late times (Fig. 1a), and the transition from initial fast exponential growth to a slow non-exponential level, and even an effective complete stop of the growth in all countries, although *none* of them have infection levels even close to those required for herd-immunity. In fact, all countries show far lower infection levels than would be expected from a simple exponential growth expected from Eq. 7 (Fig. 6).

Given some initial number of LVS-infected people (I_L^0) in the population at the point when the HVS begins to spread (i.e. the time of first HVS-infection), in a given country, one can find specific relations between the maximal number of infected people, the doubling times T_L and T_H of the two viruses, and the initial fraction of LVS and HVS infected people in the population. During the initial free spread of the LVS, it can spread until it achieves herd-immunity once the majority of the population is infected (we will take the actual population size for simplicity). The time it takes to the maximal infection point is therefore the number of doublings of the infected population from its initial value up to approximately the size of the population $\sim N$, times the doubling time

$$t_{max} = T_L \cdot \log_2 \left(\frac{N}{I_L^0} \right) \quad (10)$$

Until the LVS achieves herd-immunity, the HVS spreads freely. Its spread should follow an exponential growth until slowed/stopped once herd immunity is achieved. As discussed above, if the LVS infects the majority of the population and achieves herd-immunity before being outrun by the spread of the HVS, the mutual immunity insures that the further spread of the HVS will also be quenched. At this point the HVS would also stop and achieve its maximum infection level. Hence, the number of HVS-infected people is just

$$I_{H(t_{max})} = I_H^{max} = 2^{t_{max}/T_H} = 2^{\left(\frac{T_L}{T_H}\right) \log_2 \left(\frac{N}{I_L^0}\right)} = \left(\frac{N}{I_L^0}\right)^{T_L/T_H}, \quad (11)$$

or writing it in logarithmic terms

$$\log_2(I_H^{max}) = -\left(\frac{T_L}{T_H}\right) \log_2 \left(\frac{I_L^0}{N}\right). \quad (12)$$

The exposure time to a virus for a given country was derived in Eq. 8. It is valid both for the LVS and the HVS, where the parameters correspond to the specific strain (i.e. t_{exp}^L , t_0^L and T_L for the LVS, and t_{exp}^H , t_0^H and T_H for the HVS, and c_{exp} is the same for both (neglecting seasonal variations in mobility). Using this relation, we can find the infection level of the LVS at a given country at the point when the HVS first arrives, I_L^0 . We get

$$t_{exp}^L - t_0^L = -T_L \ln_2 c, (\text{country})$$

$$t_{exp}^H - t_0^H = -T_H \ln_2 c, (\text{country})$$

and therefore

$$t_{exp}^H - t_{exp}^L = (t_0^H - t_0^L) - (T_H - T_L) \ln_2(c_{exp}). \quad (13)$$

Plugging this into Eq. 11 we get

$$I_L^0 = 2^{(t_{exp}^H - t_{exp}^L)/T_L} = 2^{[t_0^H - t_0^L - (T_H - T_L) \ln_2(c_{exp})]/T_L} = 2^{(t_0^H - t_0^L)/T_L} c_{exp}^{1 - T_H/T_L},$$

to finally get from Eq. 12

$$I_H^{max} = \left(\frac{N}{I_L^0}\right)^{T_L/T_H} = \left(\frac{N}{2^{(t_0^H - t_0^L)/T_L} c_{exp}^{1 - T_H/T_L}}\right)^{T_L/T_H},$$

or in logarithmic terms

$$\log_2(I_H^{max}) = [(t_0^L - t_0^H)/T_H] + \left(\frac{T_L}{T_H}\right) \log_2 N - \left(\frac{T_L}{T_H} - 1\right) \log_2 c_{exp}, \quad (14)$$

i.e. the *maximal* number of infected people in a given country in which the LVS reached herd-immunity before the HVS, should be directly correlated with the total population in the country, and inversely correlated with the exposure level to China. The exact powers are determined by the doubling times (spread rates) of the LVS and the HVS. The larger the exposure to China the longer the delay between the initial LVS infection and the initial HVS infection (see Eq. 13), allowing for the LVS more time to spread and achieve herd-immunity while the HVS is still at earlier stages. The smaller the population size in a given country, the faster the LVS can achieve herd-immunity, and quench the spread of the LVS at earlier times, leading to lower infection levels, consistent with the models in Fig. 2. A critical transition can occur for less exposed and/or more populated countries, when I_H^{max} is comparable with the overall population, i.e. providing no constraint on the HVS spread (needless to say I_H^{max} needs always to be smaller or equal to the total population N). This happens when the HVS had sufficient time to spread and

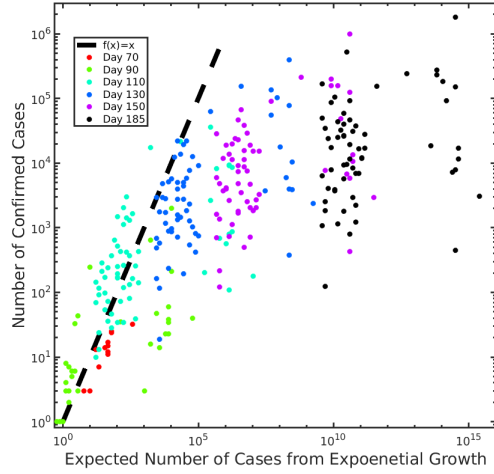


Figure 6: The number of confirmed cases vs. the expected number of cases for an exponential growth, at epochs. Early on (day 70) all infected countries followed a simple exponential growth (compare with the dashed line). At later times newly infected countries follow an exponential growth, while countries infected earlier stop following an exponential growth and show a sub-exponential growth. This patterns repeats, with more countries stopping their exponential growth after their initial exponential spread. By day 185 all countries in our sample already stopped spreading exponentially. This dynamics follows the expectation from the double-strain co-infection model (see text). The dashed line corresponds to a simple exponential growth with a doubling time of 2.7 days.

eventually outrun the spread of the LVS, before the latter achieved herd-immunity level. As discussed in the main text, we can therefore fit our model using the data on the infection-level (testing-normalized cases or age-normalized deaths) and the mobility-exposure data, to find the model parameters. In particular, we can infer the doubling-time of the LVS using a linear-regression analysis, and then also infer the delay between the onset of the LVS and the HVS (see main text), which then also provides the mobility-exposure calibration unit α mentioned above.

For any given country we can now predict the expected infection levels. Our models in Figs. 2 suggest that the HVS spread should already slow down when the infection level reaches $\sim 10\%$ of the maximal level, and that the rise and decay of the HVS infection level should be asymmetric with a fast rise and a slow decay, in particular it is more asymmetric than would be expected for the case of a single virus evolution (e.g. compare the more symmetric structure of the LVS evolution in comparison with the HVS), potentially consistent with observed asymmetries in the epidemic dynamics observed in many countries.

1.5 Dynamics of early spread

If allowed to spread with no constraint the number of infected people should rise exponentially to be

$$\log_2 [I_{\text{exp}}(t)] = (t - t_{\text{exp}})/T, \quad (15)$$

at time t , where I is the number of infected people, t_0 is the time of the initial exposure (first infection), and T is the doubling time for the the given virus. As discussed above (section 1.4), in the early stages of the HVS infection the evolution should effectively be indistinguishable from that expected for a the unconstrained spread of a single virus. In Fig. 6 we show the number of age-corrected deaths in different countries at several points in time, as a function of the initial infection time in each country. As expected, in countries infected earlier (larger number of days since first infection), the number of age-normalized deaths is larger, and consistent with an exponential growth as featured in Eq. 15. However, at later times one can observe that the number of deaths stops growing exponentially, and the overall structure of the number of deaths in each country hardly evolves (on logarithmic scales; one can still see a very slow, non-exponential growth). This behavior is explained by the double-strain co-infection model, which suggests (see Fig. 2) that the exponential growth should begin to slow down once the LVS-infection levels reach a large fraction of the population, and later become sub-exponential until it eventually stops growing.

References

- [1] Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
- [2] Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).
- [3] Haider, N. *et al.* Passengers' destinations from China: low risk of Novel Coronavirus (2019-nCoV) transmission into Africa and South America. *Epidemiol. Infect.* **148**, e41 (2020).
- [4] Van den Broeck, W. *et al.* The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infect. Dis.* **11**, 37 (2011).
- [5] Pastore y Piontti, A., Perra, N., Rossi, L., Samay, N. & Vespignani, A. *Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age* (2019).
- [6] Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science* **368**, 395–400 (2020). URL <https://science.sciencemag.org/content/368/6489/395>. <https://science.sciencemag.org/content/368/6489/395.full.pdf>.
- [7] Dietz, K. Epidemiologic interference of virus populations. *J Math Biol* **8**, 291–300 (1979).
- [8] Castillo-Chavez, C., Hethcote, H. W., Andreasen, V., Levin, S. A. & Liu, W. M. Epidemiological models with age structure, proportionate mixing, and cross-immunity. *J Math Biol* **27**, 233–258 (1989).
- [9] Adler, F. R. & Brunet, R. C. The dynamics of simultaneous infections with altered susceptibilities. *Theor Popul Biol* **40**, 369–410 (1991).
- [10] Balmer, O. & Tanner, M. Prevalence and implications of multiple-strain infections. *Lancet Infect Dis* **11**, 868–878 (2011).
- [11] Susi, H., Barr's, B., Vale, P. F. & Laine, A. L. Co-infection alters population dynamics of infectious disease. *Nat Commun* **6**, 5975 (2015).
- [12] Kucharski, A. J., Andreasen, V. & Gog, J. R. Capturing the dynamics of pathogens with many strains. *J Math Biol* **72**, 1–24 (2016).
- [13] Nickbakhsh, S. *et al.* Modelling the impact of co-circulating low pathogenic avian influenza viruses on epidemics of highly pathogenic avian influenza in poultry. *Epidemics* **17**, 27–34 (2016).
- [14] Thompson, R., Thompson, C., Pelerman, O., Gupta, S. & Obolski, U. Increased frequency of travel in the presence of cross-immunity may act to decrease the chance of a global pandemic. *Phil. Trans. R. Soc. B* **374**, 20180274 (2019). URL <https://www.biorxiv.org/content/early/2019/01/18/404871>. <https://www.biorxiv.org/content/early/2019/01/18/404871.full.pdf>.
- [15] Kamikubo, Y. & Takahashi, A. Epidemiological tools that predict partial herd immunity to sars coronavirus 2. *medRxiv* (2020). URL <https://www.medrxiv.org/content/early/2020/03/27/2020.03.25.20043679>. <https://www.medrxiv.org/content/early/2020/03/27/2020.03.25.20043679.full.pdf>.
- [16] Qi, W. *et al.* Emergence and Adaptation of a Novel Highly Pathogenic H7N9 Influenza Virus in Birds and Humans from a 2013 Human-Infecting Low-Pathogenic Ancestor. *J. Virol.* **92** (2018).
- [17] Seo, S. H. & Webster, R. G. Cross-reactive, cell-mediated immunity and protection of chickens from lethal H5N1 influenza virus infection in Hong Kong poultry markets. *J. Virol.* **75**, 2516–2525 (2001).
- [18] Khalenkov, A., Perk, S., Panshin, A., Golender, N. & Webster, R. G. Modulation of the severity of highly pathogenic H5N1 influenza in chickens previously inoculated with Israeli H9N2 influenza viruses. *Virology* **383**, 32–38 (2009).
- [19] Nfon, C. *et al.* Prior infection of chickens with H1N1 or H1N2 avian influenza elicits partial heterologous protection against highly pathogenic H5N1. *PLoS ONE* **7**, e51933 (2012).
- [20] Li, X. *et al.* Transmission dynamics and evolutionary history of 2019-nCoV. *J. Med. Virol.* **92**, 501–511 (2020).

- [21] Su, Y. C. F. *et al.* Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nat Commun* **6**, 7952 (2015).
- [22] Ferguson, N. M., Galvani, A. P. & Bush, R. M. Ecological and immunological determinants of influenza evolution. *Nature* **422**, 428–433 (2003).
- [23] Nickbakhsh, S. *et al.* Virus-virus interactions impact the population dynamics of influenza and the common cold. *Proc. Natl. Acad. Sci. U.S.A.* (2019).
- [24] Chen, Z.-w., Li, Z., Li, H., Ren, H. & Hu, P. Global genetic diversity patterns and transmissions of sars-cov-2. *medRxiv* (2020). URL <https://www.medrxiv.org/content/early/2020/05/08/2020.05.05.20091413>. <https://www.medrxiv.org/content/early/2020/05/08/2020.05.05.20091413.full.pdf>.
- [25] Yao, H. *et al.* Patient-derived mutations impact pathogenicity of sars-cov-2. *medRxiv* (2020). URL <https://www.medrxiv.org/content/early/2020/04/22/2020.04.14.20060160>. <https://www.medrxiv.org/content/early/2020/04/22/2020.04.14.20060160.full.pdf>.
- [26] Urbanowicz, R. A. *et al.* Human Adaptation of Ebola Virus during the West African Outbreak. *Cell* **167**, 1079–1087 (2016).
- [27] Diehl, W. E. *et al.* Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013-2016 Epidemic. *Cell* **167**, 1088–1098 (2016).
- [28] Hethcote, H. W. The mathematics of infectious diseases. *SIAM Review* **42**, 599–653 (2000).
- [29] MATLAB. *version 7.10.0 (R2010a)* (The MathWorks Inc., Natick, Massachusetts, 2010).
- [30] Kermack, W. O., McKendrick, A. G., Kermack, W. O. & McKendrick, A. G. Contributions to the mathematical theory of epidemics–I. 1927. *Bull. Math. Biol.* **53**, 33–55 (1991).
- [31] Britton, T., Trapman, P. & Ball, F. G. The disease-induced herd immunity level for covid-19 is substantially lower than the classical herd immunity level. *medRxiv* (2020). URL <https://www.medrxiv.org/content/early/2020/05/14/2020.05.06.20093336>. <https://www.medrxiv.org/content/early/2020/05/14/2020.05.06.20093336.full.pdf>.
- [32] Ioannidis, J. The infection fatality rate of covid-19 inferred from seroprevalence data. *medRxiv* (2020). URL <https://www.medrxiv.org/content/early/2020/05/19/2020.05.13.20101253>. <https://www.medrxiv.org/content/early/2020/05/19/2020.05.13.20101253.full.pdf>.
- [33] Havers, F. P. *et al.* Seroprevalence of antibodies to sars-cov-2 in six sites in the united states, march 23-may 3, 2020. *medRxiv* (2020). URL <https://www.medrxiv.org/content/early/2020/06/26/2020.06.25.20140384>. <https://www.medrxiv.org/content/early/2020/06/26/2020.06.25.20140384.full.pdf>.
- [34] Stephenson, I. *et al.* Cross-reactivity to highly pathogenic avian influenza H5N1 viruses after vaccination with nonadjuvanted and MF59-adjuvanted influenza A/Duck/Singapore/97 (H5N3) vaccine: a potential priming strategy. *J. Infect. Dis.* **191**, 1210–1215 (2005).
- [35] Manenti, A. *et al.* Evaluation of SARS-CoV-2 neutralizing antibodies using a CPE-based colorimetric live virus micro-neutralization assay in human serum samples. *J. Med. Virol.* (2020).
- [36] Sekine, T. *et al.* Robust t cell immunity in convalescent individuals with asymptomatic or mild covid-19. *bioRxiv* (2020). URL <https://www.biorxiv.org/content/early/2020/06/29/2020.06.29.174888>. <https://www.biorxiv.org/content/early/2020/06/29/2020.06.29.174888.full.pdf>.
- [37] Fongaro, G. *et al.* Sars-cov-2 in human sewage in santa catalina, brazil, november 2019. *medRxiv* (2020). URL <https://www.medrxiv.org/content/early/2020/06/29/2020.06.26.20140731>. <https://www.medrxiv.org/content/early/2020/06/29/2020.06.26.20140731.full.pdf>.
- [38] Huang, A. T. *et al.* A systematic review of antibody mediated immunity to coronaviruses: antibody kinetics, correlates of protection, and association of antibody responses with severity of disease. *medRxiv* (2020).
- [39] Hale, N., T., Angrist, B. K., Petherick, A., Phillips, T. & Webster, S. Variations in government responses to covid-19. *Blavatnik School of Government, University of Oxford* **BSG-WP-2020/032** (2020).