covid19.Explorer: A web application and R package to explore United States COVID-19 data

Liam J. Revell[1,2]

[1]**Department of Biology, University of Massachusetts Boston, Boston, Massachusetts 02125**
[2]**Facultad de Ciencias, Universidad Católica de la Santísima Concepción, Concepción, Chile**

Corresponding author:
Liam J. Revell[1,2]

Email address: liam.revell@umb.edu

ABSTRACT

Beginning at the end of 2019, a novel virus (later identified as SARS-CoV-2) was characterized in the city of Wuhan in Hubei Province, China. As of the time of writing, the disease caused by this virus (known as COVID-19) has already resulted in over 2 million deaths worldwide. SARS-CoV-2 infections and deaths, however, have been highly unevenly distributed among age groups, sexes, countries, and jurisdictions over the course of the pandemic. Herein, I present a tool (the *covid19.Explorer* R package and web application) that has been designed to explore and analyze publicly available United States COVID-19 infection and death data from the 2020/21 U.S. SARS-CoV-2 pandemic. The analyses and visualizations that this R package and web application facilitate can help users better comprehend the geographic progress of the pandemic, the effectiveness of non-pharmaceutical interventions (such as lockdowns and other measures, which have varied widely among U.S. states), and the relative risks posed by COVID-19 to different age groups within the U.S. population. The end result, I hope, is an interactive tool that will help its users develop a improved understanding of the temporal and geographic dynamics of the SARS-CoV-2 pandemic, accessible to lay people and scientists alike.

INTRODUCTION

In 2019, a novel infectious disease was first identified in Wuhan, a city of approximately 11 million residents located in the Hubei Province of China. This infectious disease, called *Coronavirus disease 2019*, or COVID-19 (Velavan and Meyer, 2020), is now known to be caused by the previously unidentified *severe acute respiratory syndrome coronavirus 2* or SARS-CoV-2 (Wong et al., 2020). Following the Wuhan outbreak, cases of SARS-CoV-2 infection and death were subsequently identified in Europe, the United States, and (by the time of writing) at least 192 countries worldwide. Counting from the beginning of this global pandemic, there have been over 2.3 million confirmed COVID-19 deaths, more than 400,000 of which have occurred in the United States alone (CDC, 2021).

R (R Core Team, 2020) is a powerful scientific computing environment used by statisticians, data scientists, academic researchers, and students worldwide. I have built a multifunctional R package (*covid19.Explorer*) and corresponding web application (https://covid19-explorer.org). The purpose of both is to aid both scientists and lay people better understand the 2020 COVID-19 pandemic in the United States. Although my focus is on U.S. COVID-19 data, readers from other countries might also be interested in the project – for instance, because the seasonal dynamics of infection or the age distribution of mortality has been broadly similar among different affected areas of the globe.

1 METHODS

1.1 Preamble

covid19.Explorer is a library of functions and data that can be loaded and run using the R scientific computing software (R Core Team, 2020). The *covid19.Explorer* package is open source and freely available from its GitHub page (https://github.com/liamrevell/covid19.Explorer/). The *covid19.Explorer* package in turn depends on the CRAN R packages *maps* (Becker et al., 2018), *phytools* (Revell, 2012), *randomcoloR* (Ammar, 2019), and *RColorBrewer* (Neuwirth, 2014).

Though the *covid19.Explorer* R package can be downloaded, installed, and run from R on its own, it has primarily been designed to be utilized via a web portal: https://covid19-explorer.org. This web portal was built in the integrated development environment *Rstudio* (RStudio Team, 2020), using the web application development system *shiny* (Chang et al., 2021). In addition to those R libraries already mentioned, the web application also uses the package *shinyWidgets* (Perrier et al., 2021).

The data used by the various applications of the *covid19.Explorer* are all publicly available and were obtained (unless otherwise indicated) from the *United States Centers for Disease Control and Prevention National Center for Health Statistics* (https://www.cdc.gov/nchs/) or the *United States Census Bureau* (https://www.census.gov/).

1.2 Types of functions in *covid19.Explorer*

The *covid19.Explorer* R package (and corresponding web application) consists of two main types of functions.

The first of these (exemplified by the *shiny* web application webpage tabs denominated *U.S. COVID-19 infections*, *Iceberg plot*, *State comparison*, *Plausible range*, and *Infection estimator*) consists of functions that are designed to estimate the true number of COVID-19 infections over the course of the pandemic. For reasons that I will elaborate below, confirmed COVID-19 cases underestimate the true number of infections, usually by a substantial margin (Wu et al., 2020). Since there are a variety of reasons that the true number of infections (rather than simply the number of confirmed cases) is of interest, these various aforementioned applications of the *covid19.Explorer* package are all designed to help users apply a model (of their own design, see below) to estimate the daily number of new infections, the plausible range of new infections, the cumulative number of infections, or the daily or cumulative infections as a percentage of the total population or per 1M persons.

Each of these *covid19.Explorer* applications uses a model – but it is one whose parameters must be set by the user. That is to say, users of the *covid19.Explorer* package or corresponding web interface will need to specify: (1) a value or set of values for the infection fatality ratio, IFR (Roques et al., 2020), of SARS-CoV-2 infection through time; and (2) an average lag time from infection to death. The values for each of these model parameters have been default values that are fairly reasonable, as detailed in the sections below; however, users are nonetheless strongly encouraged to apply multiple values and examine the sensitivity of their results. (In fact, this is one of the main purposes of the project!)

The second type of function (exemplified by the web application tabs *Deaths by age*, *Excess mortality by age*, and *By state*) *do not* employ an explicit model and exist primarily to permit the user to interact directly with U.S. *Centers for Disease Control* (CDC) COVID-19 death and 2020 excess mortality data, to understand the implications of these data, and to generate interesting or useful data visualizations.

The names and corresponding web application tabs of *all* functions in *covid19.Explorer* are given in alphabetical order in Table 1, below.

1.3 Estimating infections

Since the beginning of this pandemic, it has been widely understood that confirmed COVID-19 cases underestimate the true number of infections, sometimes vastly (Al-Sadeq and Nasrallah, 2020; Wu et al., 2020). This is likely to be due to multiple factors. One important factor is that there has been limited testing capacity throughout much of the COVID-19 pandemic in the United States, but particularly when the pandemic was in its earliest days (Rosenberg et al., 2020). A second significant factor affecting the disconnect between observed cases and true infections are the facts that in the United States SARS-CoV-2 testing is voluntary, population surveillance testing has been relatively scarce, and many cases of SARS-CoV-2 infection present asymptotically or with mild symptoms (Oran and Topol, 2020). As such, I consider confirmed COVID-19 deaths to be a much more reliable indicator of disease burden than confirmed cases. Deaths, however, are obviously a *lagging* indicator of infections.

Table 1. A summary of the functions and corresponding web applications that currently make up the *covid19.Explorer* R package.

Function name	Application tab	Description
age.deaths	Excess mortality by age	Graph weekly or cumulative excess mortality by age and jurisdiction.
compare.infections	State comparison	Compare daily or cumulative deaths and estimated daily or cumulative infections between states and U.S. jurisdictions.
covid.deaths	Deaths by age	Plot weekly or cumulative confirmed COVID-19 deaths by age group and compared to all deaths.
iceberg.plot	Iceberg plot	Graph observed daily confirmed COVID-19 cases (above the 'waterline' of the graph and estimated unobserved infections (below it).
infection.estimator	Infection estimator	Estimate daily or cumulative COVID-19 infections based on observed deaths and confirmed cases.
infection.range.estimator	Plausible range	Estimate the plausible range of daily or cumulative infections based on an interval of IFR values at each time point.
infections.by.state	COVID-19 infections	Visualize geographic distribution of new or cumulative COVID-19 infections through time.
state.deaths	By state	Graph weekly or cumulative excess deaths by U.S. state.

Consequently, I chose to use a model in which the number of new infections on the ith day was estimated by simply taking the number of observed COVID-19 deaths on day $i + k$ (in which k is the average lag period between initial infection and death, where death is the outcome of infection), and then dividing this quantity by the infection fatality rate (also called the *infection fatality ratio* or *IFR*). In other words, given 50 COVID-19 deaths on day $i + k$, and an IFR of 0.5%, we would predict that 10,000 new SARS-CoV-2 infections *had occurred* on day i. Both k, the average lag time from infection to death (in cases of SARS-CoV-2 infections resulting in death), and the IFR are to be specified by the user.

A fairly reasonable lag time between infection and death might be approximately three weeks. For example, during a large outbreak in Melbourne, Australia the time difference between the peak recorded cases and peak confirmed COVID-19 deaths was around 17 days. Infected persons normally test negative for the first few days following exposure (Kucirka et al., 2020), so this more or less corresponds with a lag-time of three weeks.

Likewise, IFR values ranging from about 0.3% to over 1.0% have been reported over the course of the pandemic. For instance, a study based on an early, super-spreader event in Germany estimated an IFR (corrected to the demographic distribution of the local population) of 0.36% (Streeck et al., 2020). Other researchers have reported higher estimated IFR. For example, an Italian study estimated an overall infection fatality rate of 1.31% (Rinaldi and Paradisi, 2020). In general, it is probably reasonable to suppose that IFR has fallen through time as treatment of severely ill patients has improved (Fan et al., 2020). Likewise, even within the U.S., IFR is unlikely to be precisely the same at a given date in different jurisdictions, due to differences in demographic structure between areas.

I suspect that it is within reason for users of *covid19.Explorer* to specify an IFR that is no greater than about 1.0% and that declines gradually from the start of the pandemic towards the present, with a current IFR that is perhaps around 0.3% - 0.5%. *covid19.Explorer* permits the user to specify a time-varying IFR by fixing the IFR at each quarter (on the website), or at any arbitrary time interval (using the R package directly), and then interpolating daily IFR between each period using local regression smoothing (LOESS; Cleveland 1979).

Reporting can vary through time including regularly over the course of the week. (For instance, fewer COVID-19 deaths tend to be reported on the weekends compared to Monday through Friday.) To

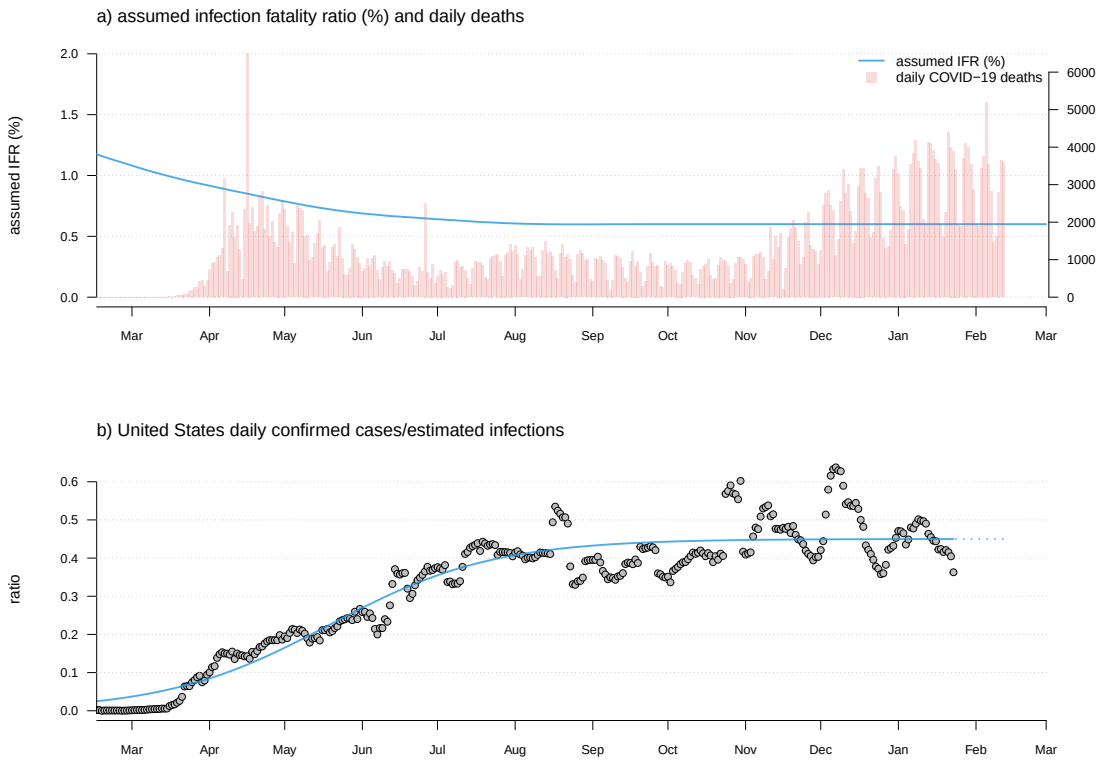

Figure 1. a) Observed U.S. daily COVID-19 deaths (red bars) and assumed IFR function (blue line). b) Ratio of confirmed COVID-19 cases / estimated infections.

take these reporting artifacts into account we used both moving averages and local regression (LOESS) smoothing. Both the window for the moving average and the LOESS smoothing parameter are controlled by the user.

The approach of using only confirmed COVID-19 deaths – though robust – does not permit us to estimate the true number of infections between k days ago and the present. For this, I have assumed a sigmoidal relationship between time and the ratio of confirmed cases over the true number of infections (Figure 1). Since the true number of new infections cannot exceed the number of confirmed cases, logic dictates that this ratio must have a value between 0 and 1.

I decided on a sigmoidal relationship because it seemed reasonable to assume the ratio was very low early in the pandemic when confirming a new infection was limited primarily by testing capacity, but has probably risen (in many localities) to a more or less consistent value as testing capacity increased. Since getting tested is voluntary, and since many infections of SARS-CoV-2 are asymptomatic or only mildly symptomatic, this ratio seems unlikely to rise to very near 1.0 in the U.S. regardless of testing.

Figure 1, created using *covid19.Explorer*, shows daily estimated infections (under our model) / confirmed cases for all U.S. data over the entire course of the pandemic to date in panel Figure 1b, given observed daily deaths (red bars) and assumed IFR evolution (blue curved line) of panel Figure 1a.

Our plot seems to show a ratio of confirmed / estimated infections of about 0.45 at the present; however, the reader should keep in mind that in practice this value is estimated separately for each jurisdiction that is being analyzed, and as such might be lower in some states and higher in others, even for a constant IFR value or function.

After fitting this sigmoidal curve to our observed and estimated cases through now $-k$ days, we then turn to the last period. To obtain estimated infections for these days, we simply divide our observed cases from the last k days of data by our fitted values from the sigmoid curve. Figure 2 shows the result of this analysis applied to data for the U.S. state of Massachusetts.

In addition to computing the raw number of daily infections, this method can also be used to compute

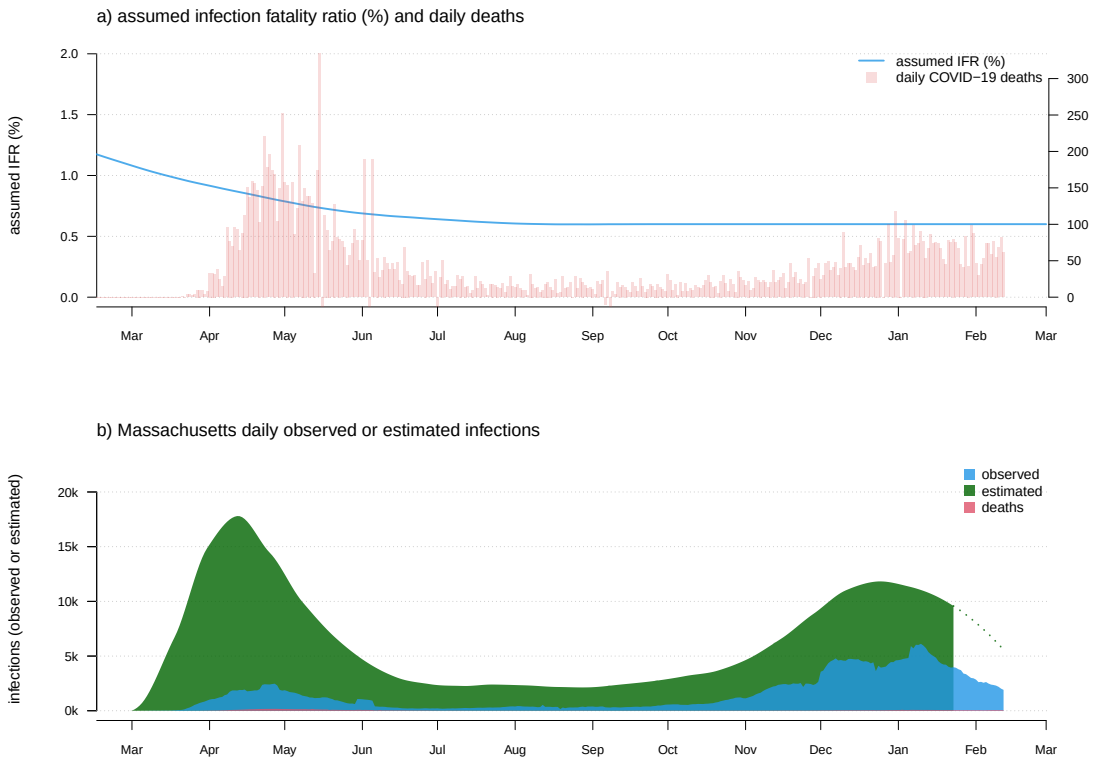

Figure 2. a) Observed daily COVID-19 deaths and an assumed model of IFR in which the infection fatality ratio is initially high (1.5%), but then declines and stabilizes at around 0.6% through the present day. b) Estimated daily infections (green), cases (blue), and deaths (red).

infections as a percentage of the total population. To make this calculation, I obtained state populations through time from the U.S. Census Bureau. Data is only given through 2019, so to estimate state-level 2020 population sizes, we used a total mid-year 2020 U.S. population estimate of 331,002,651) to 'correct' each 2019 state population size to a 2020 level.

Finally, CDC mortality data splits New York City (NYC) from the rest of New York state. Since this contrast is interesting (e.g., Gonzalez-Reiche et al. 2020), I maintained the separation – and used a mid-2019 population estimate of (8,336,817) for NYC, then simply assumed that the population of NYC has changed between 2015 and 2020 in proportion to the rest of the state. (Since they have a part : whole relationship, this seemed pretty reasonable.)

1.4 Assumptions about estimating infections

This model is very simple. In using it, we merely start by imagining that if we knew the true number of infections and the IFR for our population of interest on day i, then we could predict the number of deaths on day $i + k$, in which k is the lag-time from infection to death (for SARS-CoV-2 infections leading to death). Having observed the deaths, and supposing a particular value of IFR for day i, we can likewise work backwards and reconstruct the most plausible number of infections for that day.

Although the model does not pre-suppose a specific value or function for IFR, it does require that one be specified by the user. As such, it is probably worth mentioning the effect of setting an IFR value that is too high or too low compared to the (invariably unknown) *true* IFR for the population of interest. An IFR that is too high (overall or at a specific time during the pandemic) will have the general effect of causing us to systematically underestimate the number of infections that have occurred. This makes sense because if we imagine observing 50 COVID-19 deaths, an IFR of 0.5% would imply that these deaths correspond to a total of 10,000 infections. By contrast, a higher IFR of, say, 1.0% would instead imply that only 5,000 infections had occurred. Assuming an IFR value that is too low will (obviously) have

exactly the opposite effect and thus cause us to overestimate the number of infections that have occurred.

We assume a homogeneous value of k at any particular time. In fact, literature sources report lag-times between two and eight weeks (e.g., Yang et al. 2020). Nonetheless, we assert that inferences by our method should not be badly off – so long as IFR does not swing about wildly from day to day, and so long as the number of deaths is not extremely few for any reporting period. We likewise assume a constant lag-period, k, through time. This assumption is perhaps a bit more dubious as it seems quite reasonable to suppose that for a specific state or jurisdiction that as IFR falls, k might also increase. This is a complexity that we explicit chose to ignore in our model.

We assume that a more or less consistent fraction of COVID-19 deaths are reported as such – that is, that COVID-19 is neither systematically under- or overreported as the cause of death at any point during the course of the pandemic. A violation of this assumption is not quite as grave as it might seem, however, because it can simply be 'baked in' to our model for IFR. For instance, if we think that COVID-19 deaths were under-reported near the start of the pandemic (due to limited testing capacity), this can be accommodated into our model for daily infections simply by specifying a slightly lower IFR value at that time (keeping in mind, of course, that the true IFR has generally been falling through time; e.g., Levin et al. 2020).

In estimating the number of daily infections from k days ago to today, we assume that the relationship between time (since the first infections) and the ratio of confirmed and estimated infections is sigmoidal in shape (Figure 1). This is a testable assumption that seems to hold fairly well across the entire U.S. (Figure 1) and for some jurisdictions, but less well for others. It is equally plausible to suspect that this ratio could shift not only as a function of time, but also as demands on testing capacity rise and fall with case numbers. This should be the subject of additional study, but our suspicion is that this would not be likely to have a large effect on our model compared to other simplifications.

Finally we assume no or limited reporting delay. This is obviously incorrect. There are two main sources of reporting delay: the delay between when an individual is infected and when they go on to test positive for SARS-CoV-2; and the delay between when an infected patient dies and their death is reported to the CDC. Given this delay in reporting, a more precise interpretation of the estimated number of daily infections, is a (rough) estimate of the number of new individuals who would test positive for SARS-CoV-2 under a hypothetical scenario of universal testing.

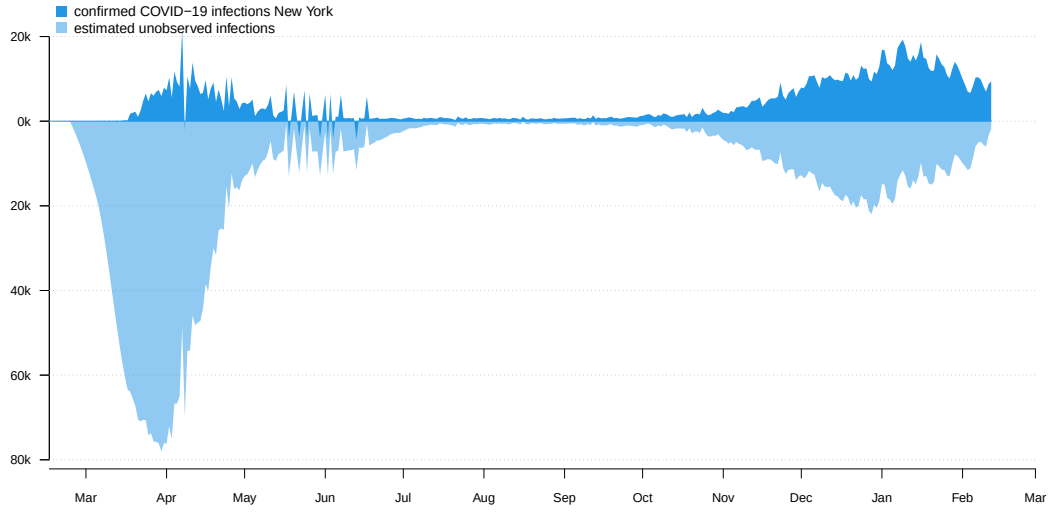

Figure 3. Iceberg plot showing the confirmed daily new infections (above the 'waterline' of the plot) and estimated unobserved infections (below it) for New York state.

1.5 Showing observed and estimated unobserved infections using an 'iceberg plot'

As noted above, it has long been well-understood that the number of daily confirmed COVID-19 cases is an underestimate of the true number of daily infections, sometimes by a very wide margin (Wu et al., 2020). To visualize this phenomenon, I devised an iceberg plot in which we simultaneously graph the number of observed COVID-19 infections (above the 'waterline' of the graph) and the estimated number of unobserved infections (below it). Figure 3 gives this analysis for New York state, in which I assumed the same IFR model through time as was used for Figures 1 and 2.

1.6 Mapping the distribution of infections across states

A hallmark feature of the U.S. COVID-19 pandemic has been the shifting geographic distribution of infections through time among states. To capture this dynamic, I devised a plotting method for *covid19.Explorer* in which I overlay the daily or cumulative COVID-19 infections under our model (outlined above), separated by state.

For this visualization, I selected a geographic color palette such that RGB color values were made to vary as a function of latitude, longitude, and (arbitrarily) geographic distance from Florida. This is intended to have the effect of making the regional geographic progression of infection more apparent in the graph. The result can be seen in Figures 4 and 5.

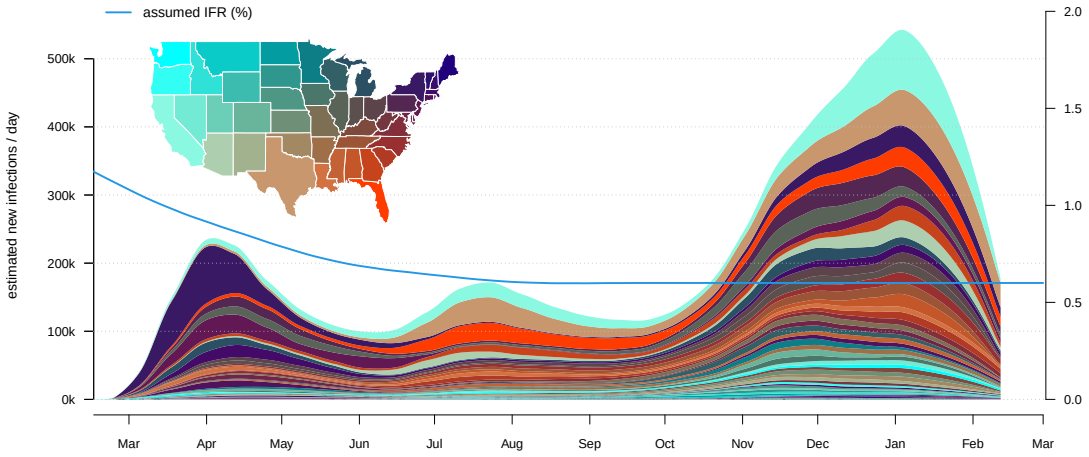

Figure 4. Daily estimated infections separated by state.

This plotting method shares all the assumptions of our infection estimator, above, but adds the additional assumption that our model of IFR is the same for all states. This assumption is quite dubious, in fact, as IFR could be expected to rise in locations were hospital resources are overtaxed by high disease burden; and, conversely, fall in hospitals were staff have more experience in treating COVID-19 patients. Although I do not doubt that these nuances are important in making specific, quantitative statements about the particular number of infections in each state, I nonetheless believe that my method is effective at visually capturing the overall geographic dynamics of the COVID-19 pandemic in the United States.

1.7 Visualizing COVID-19 mortality data

In addition to modeling the number of infections through time, the *covid19.Explorer* R package and website also allows users to visualize the distribution of COVID-19 deaths by age and sex, as well as mortality in excess of normal during 2020 compared to other recent years (2015-2019).

Excess mortality (also called *mortality displacement*; e.g., Huynen et al. 2001) is defined as the number of deaths (for any period) in excess of the 'normal' number of deaths for the same period. To compute the raw death counts for any jurisdiction, I tabulated the 2015-2018 counts with the 2019-2020 provisional counts. To correct observed deaths in prior years to 2020 levels, I simply multiplied the past-year death tally by the ratio the jurisdiction population in 2020 compared to the population in the past year. Finally, to compute excess deaths for any jurisdiction, I then took the death counts (or corrected

death counts) for 2020, and subtracted the mean of years 2015 through 2019. This treats 2015 through 2019 as 'normal' years, and 2020 as unusual.

1.8 The *covid19.Explorer* web interface

Though the *covid19.Explorer* package can be used within an interactive R session, it has also been interfaced to the web by way of the web application that I developed in Rstudio (RStudio Team, 2020) using the *shiny* web development system (Chang et al., 2021). The *covid19.Explorer* web application is hosted at the website https://covid19-explorer.org.

Figure 5 shows a screenshot of the web application, illustrating an analysis of the estimated cumulative infections through time across U.S. states. In this web application, the user must specify the value of IFR at the beginning of each three month period, and that at the end of the year, beginning on Feb. 1, 2020, and ending on Jan. 31, 2021. Values on these intervals are interpolated using LOESS smoothing.

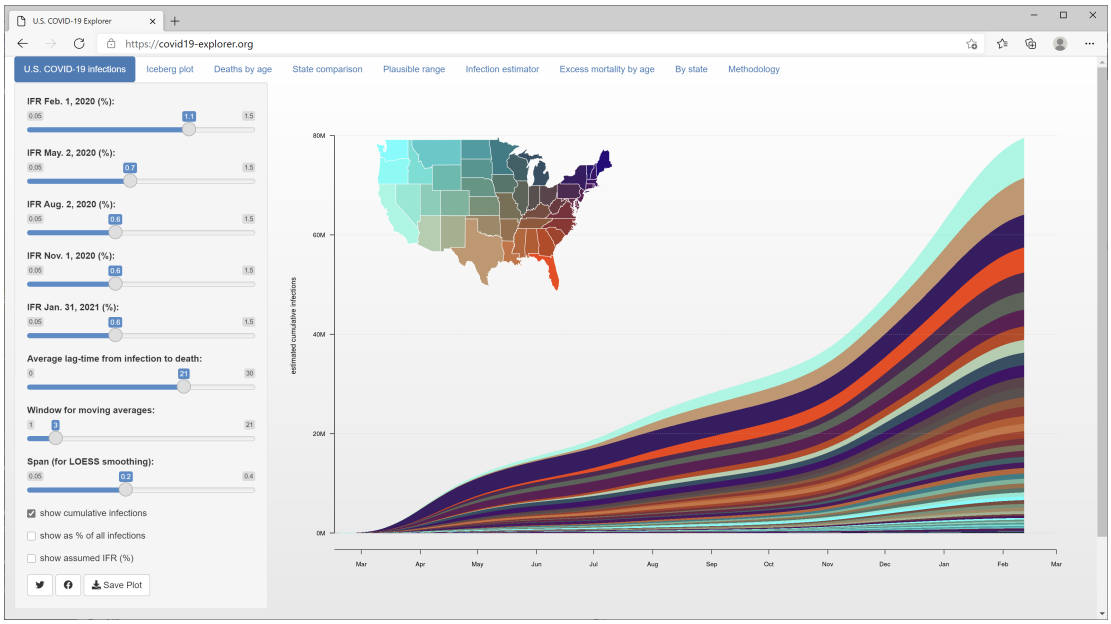

Figure 5. The *covid19.Explorer* web interface (https://covid19-explorer.org) showing estimated cumulative infections among states under the same IFR model as Figures 1 - 4.

2 RESULTS

The purpose of this article is to describe a software tool, which I have largely done in the preceding section. Here, I will attempt to highlight some results and insights that can be obtained by users via interaction with the *covid19.Explorer* R package or web application.

2.1 Herd immunity and the cumulative proportion of the population infected

The question of cumulative percent infected is relevant to the (unnecessarily controversial) concept of 'herd immunity' (Randolph and Barreiro, 2020). The herd immunity threshold (HIT), whether reached via natural infection or vaccination, is typically defined as the proportion of the population that must be immune in order to cause the basic reproductive number of the virus (R_t) to fall below 1.0, absent mitigations (Anderson and May, 1985). When R_t has fallen below 1.0, daily new infections should progressively decline.

The HIT is normally estimated by taking the reproductive number when 100% of the population is susceptible (i.e., when a new disease emerges, R_0), and computing $1 - 1/R_0$. For SARS-CoV-2 various values of R_0 have been represented in the literature, from as low as around $R_0 = 2.4$ (e.g., D'Arienzo and Coniglio 2020), to as high as about $R_0 = 5.8$ (e.g., Ke et al. 2020). A value of R_0 equal to 3.0, for example, would imply that herd immunity should be reached after $1 - 1/3$ or around 67% of the population have acquired immunity through natural infection or vaccination.

The *covid19.Explorer* R package and web application can be used to evaluate the proportion of individuals in the total population that have been potentially infected with SARS-CoV-2, given our model for COVID-19 IFR through time. The following graph (Figure 6) shows estimated infections as a fraction of the total population for the U.S. state of Texas, using the same IFR model as in Figures 1, 2, 3, and 4.

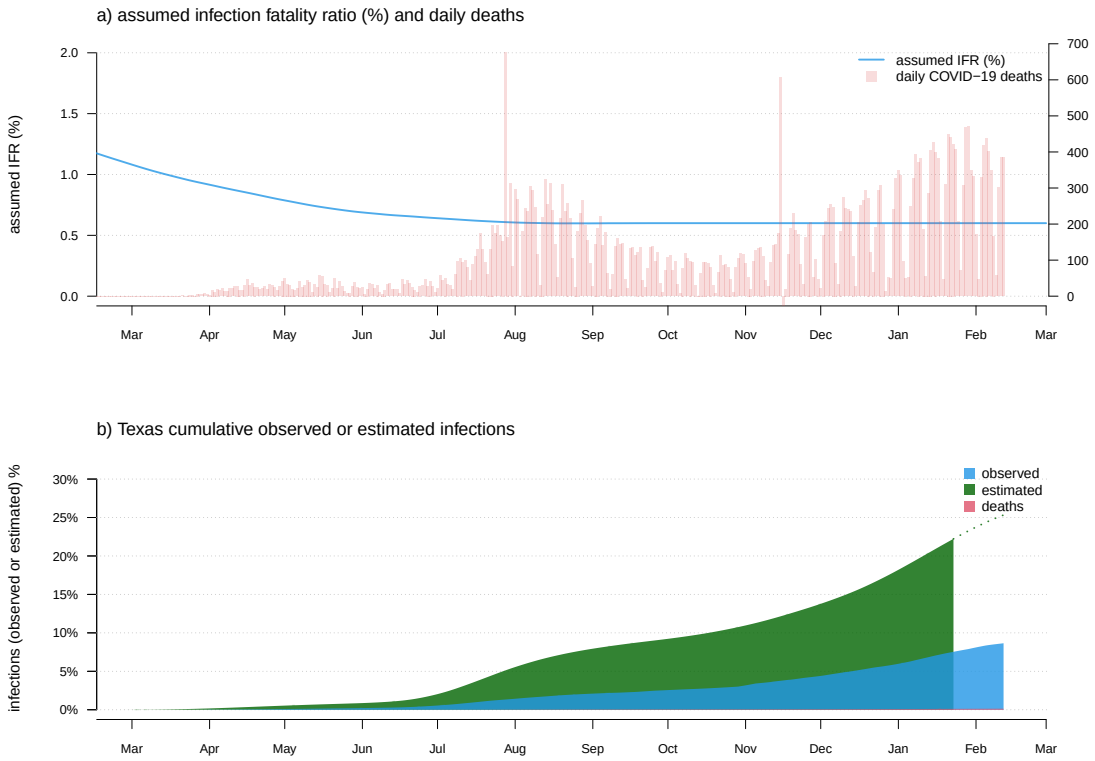

Figure 6. a) Observed daily COVID-19 deaths and an assumed model of IFR. b) Estimated cumulative infections (green), cases (blue), and deaths (red), as a percentage of the total population of the state.

Though Figure 6 suggests that perhaps around 25% of the population in Texas has already been infected, users should keep in mind that this is entirely dependent on how we decided to specify our model of IFR through time! Likewise, though this fraction is considerable, it is still well below the level of infection (e.g., 67%) required to achieve herd immunity given the majority of published estimates for R_0 of SARS-CoV-2. It may be worth noting that some authors have pointed out that the herd immunity threshold from a natural epidemic could be considerable lower than the $1 - 1/R_0$ level expected for random vaccination (e.g., Britton et al. 2020; Gomes et al. 2020). This is an intriguing possibility, and one that could be qualitatively examined with some of the tools of the *covid19.Explorer* package.

2.2 Computing a plausible range of infection numbers

A relatively simple extension of the infection estimation method, described above, is simply to admit uncertainty about the specific value of the infection fatality rate at any particular time during the pandemic, and then measure the sensitivity of our prediction to a *wide range* of different values for IFR.

This is a potentially valuable exercise, precisely because the question of the IFR for COVID-19 has been the subject of considerable controversy and confusion (e.g., Vermund and Pitzer 2020). This model can be design to accommodate an assumption of broad uncertainty in IFR early during the pandemic, with both decreasing IFR, as well as decreasing *uncertainty* in IFR, towards the present. This is illustrated for data from the U.S. state of Louisiana in Figure 7.

It should be noted that although the shaded region around the mean number of daily or cumulative infections in Figure 7 looks like a confidence band, it would only be valid to consider it as such if our high and low values of the IFR through time represented a *confidence interval* around the true infection fatality

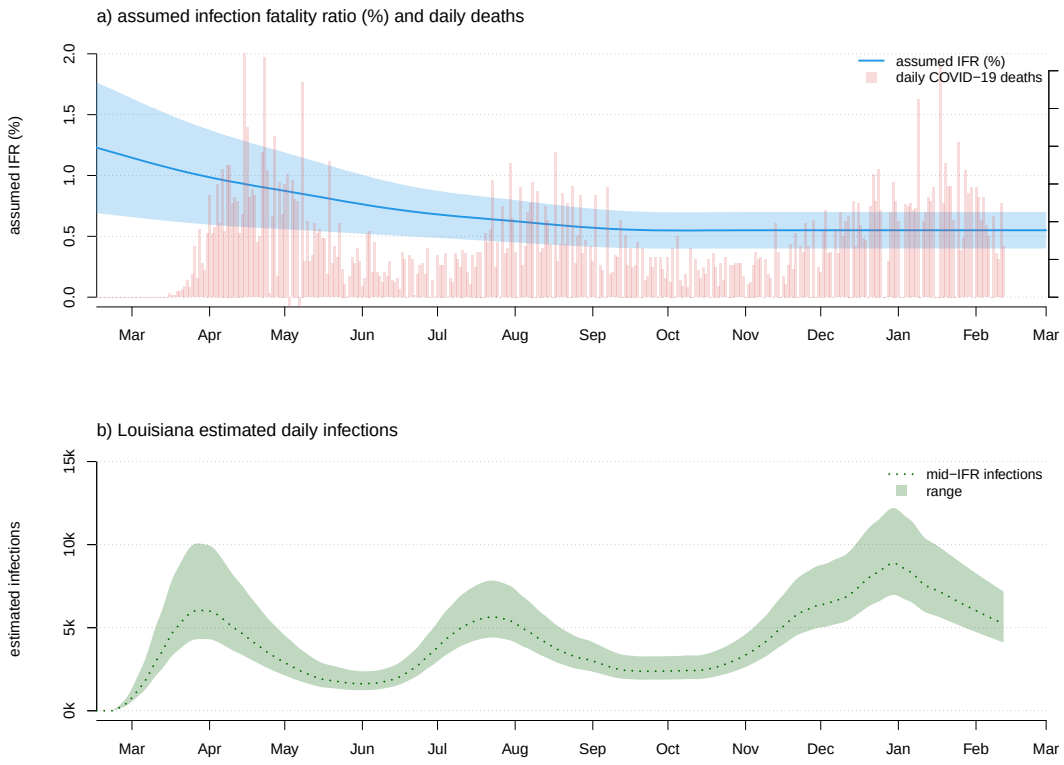

Figure 7. a) Confirmed COVID-19 deaths. b) A corresponding plausible range of daily COVID-19 infections, under our model, for the U.S. state of Louisiana.

rate (and, even then, this confidence interval would only take into account one source of uncertainty about the real daily number of infections – the IFR).

2.3 Comparing daily and cumulative infections between states

Another straightforward extension of our above-described model involves directly comparing daily (or cumulative) infections between states.

This, likewise, could be a useful activity because many readers have undoubtedly observed how common it has become for (particularly) popular press sources to attribute different infection dynamics in different states to one public health intervention or another. This attribution may be valid in many instances, but is often confounded by varying infection dynamics through time in the different states being compared. In general, evaluation of non-pharmaceutical interventions on the spread of SARS-CoV-2 (e.g., Liu et al. 2021) has been both very difficult and problematic. In Figure 8, I compare the daily confirmed deaths and estimated infections between the U.S. states of California and Florida.

This plotting method obviously shares *all* of the assumptions of our infection estimator, and (just like our method for visualizing the geographic dynamics of the pandemic across all U.S. states) requires that we use the same IFR model for each state.

Since the daily and cumulative number of infections scales with population size, valid state-to-state comparisons really only make sense if done on a per-capita basis (e.g., infections or deaths / 1M population), just as shown here in Figure 8.

2.4 COVID-19 mortality and age

Lastly, in addition to modeling the number of SARS-CoV-2 infections through time, the *covid19.Explorer* package can be used to analyze and graph COVID-19 deaths by age and sex, as well as excess mortality by age and jurisdiction.

This functionality too can sometimes lead to valuable insights.

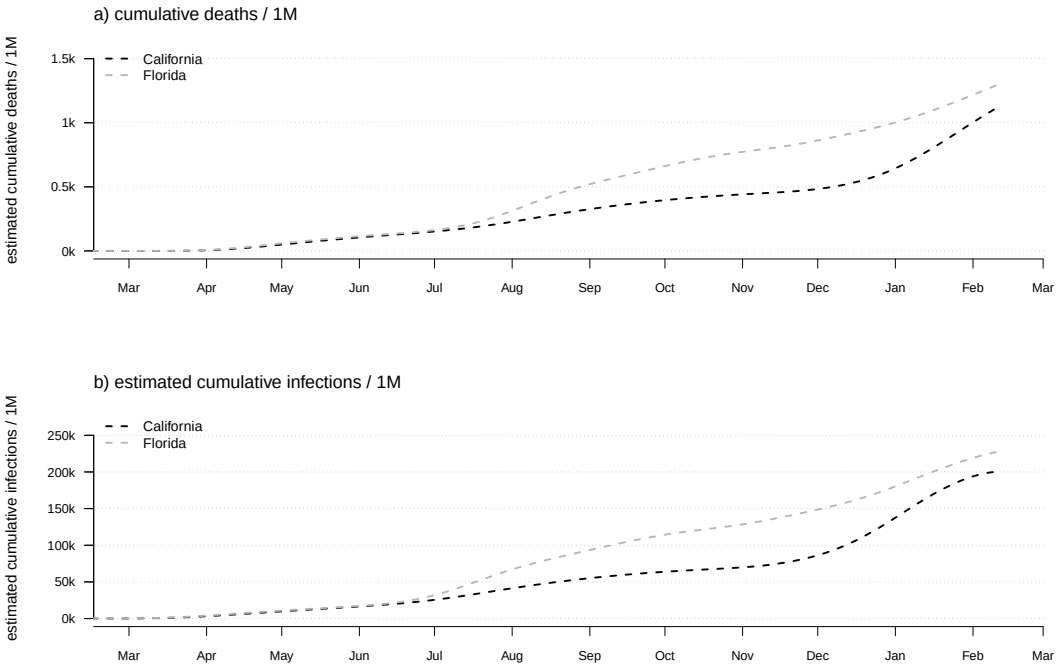

Figure 8. Daily confirmed COVID-19 deaths (a) or estimated infections (b) in the U.S. states of California vs. Florida.

For instance, it was widely predicted by media and public health experts that school and college reopening in the fall would like to increased SARS-CoV-2 infections and increased COVID-19 deaths among U.S. children and young people (e.g., Bansal et al. 2020).

In fact, CDC mortality data show exactly the opposite pattern. Figure 9 gives the weekly COVID-19 deaths overall (in panel a) and for 15-24 year olds (in panel b). We see that although the highest peaks of weekly COVID-19 deaths in the general population occurred in the Spring of 2020 and the Fall/Winter of 2020/21, peak deaths among 15-24 better align to the Summer of 2020: precisely when schools and colleges were out of session for all students.

3 DISCUSSION

The SARS-CoV-2 global pandemic of 2020 and 2021 has upended economies and civil society worldwide. With widespread vaccination campaigns underway in many countries, the pandemic may finally be in its waning days. Nonetheless, understanding the temporal and geographical dynamics of SARS-CoV-2 infections and COVID-19 deaths remains a critically important endeavor.

In this article, I present an accessible tool – the *covid19.Explorer* R package and corresponding web application – that is designed to be used to model U.S. SARS-CoV-2 infections through time, to understand the differences in epidemic dynamics between states and jurisdictions, to visualize the geographic progress of infection among U.S. states, to graph confirmed COVID-19 deaths by age and sex, and to compute and visualize excess mortality by age and jurisdiction.

Given the impact the SARS-CoV-2 pandemic has had on almost all of our daily lives over the past year, most readers of this article will know (or will be unsurprised to learn) that many other software tools and web-based applications have been developed to help visualize or better understand the temporal or geographic dynamics of COVID-19 in the United States. I nonetheless believe, however, that *covid19.Explorer* application, which has now been online (in one form or another) for nearly seven months, contains a number of different functionalities and graphics not readily available in other competing tools.

Firstly, no other software or web application, to my knowledge, lets the user build a *custom model* for the evolution of infection fatality rate (IFR) through time. This facility, offered by *covid19.Explorer*,

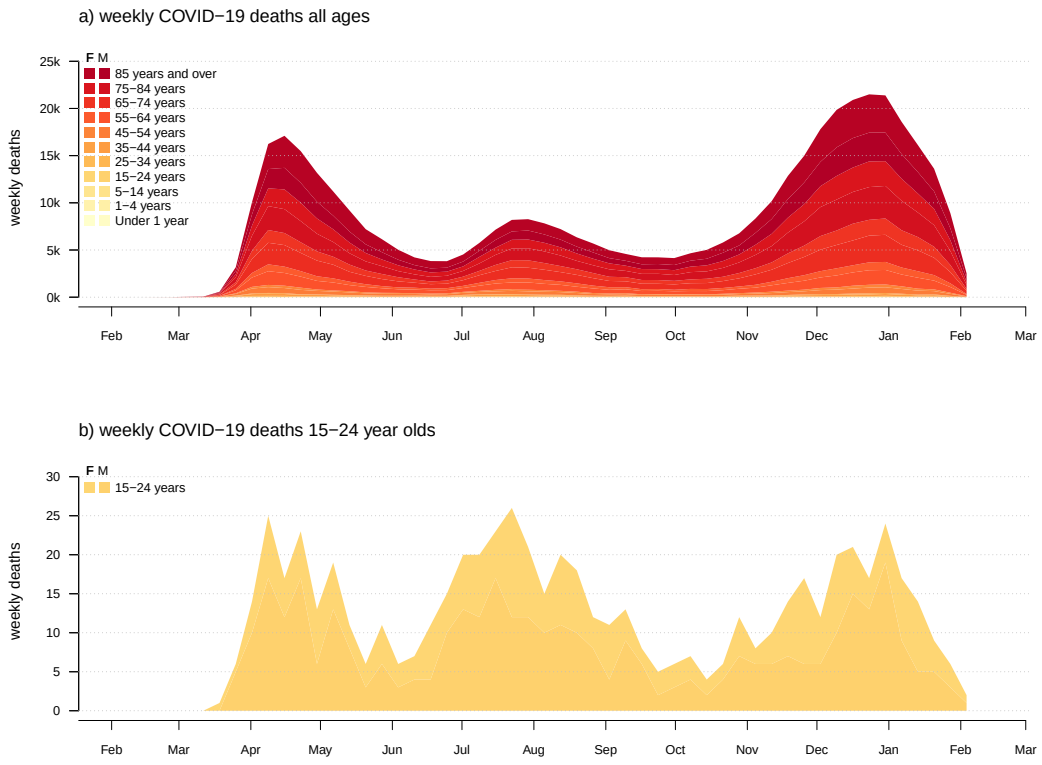

Figure 9. Weekly confirmed COVID-19 deaths for (a) all ages; and (b) individuals aged 15-24 years old in the U.S.

allows the scientists and lay people that interact with the software to design their own parameter function (based a hypothesis or external information) that will then be used to estimate infections under the model. Likewise, the tool allows *covid19.Explorer* users to progressively adjust the assumptions of this model and see how their results change in turn.

Secondly, multiple *visualization methods* of the *covid19.Explorer* R package and webpage are simply *not represented* in other software packages. For instance, I have never observed a graph similar to that of Figure 4 of this article in a publication or popular media source (other than those reporting on this application). Similarly, while it is extremely common to see graphs in the *New York Times* or other media showing the number of *confirmed* COVID-19 cases per day (the part above the 'waterline' in our iceberg graph of Figure 3), I have likewise *never* once seen a similar plot giving an estimate of the daily number of unobserved infections (below it).

Lastly, the *covid19.Explorer* package is completely transparent and open source. It pulls the data directly from public, government repositories. All model assumptions (even those not explicitly described in this paper) are readily identified from the software source code.

Even if COVID-19 soon becomes a distant memory, I hope that this tool (which I plan to make available indefinitely) will continue to be of use to scientists and educated lay people interested in the learning from the successes and failures of policy during the 2020/21 pandemic – perhaps to ensure that there are more of the former and fewer of the latter in our next global infectious disease pandemic.

ACKNOWLEDGMENTS

This research was funded in part by grants from the National Science Foundation (DBI-1759940) and FONDECYT, Chile (1201869).

REFERENCES

Al-Sadeq, D. W. and Nasrallah, G. K. (2020). The incidence of the novel coronavirus SARS-CoV-2 among asymptomatic patients: A systematic review. *International Journal of Infectious Diseases*, 98:372–380.

Ammar, R. (2019). *randomcoloR: Generate Attractive Random Colors*. R package version 1.1.0.1.

Anderson, R. M. and May, R. M. (1985). Vaccination and herd immunity to infectious diseases. *Nature*, 318(6044):323–329.

Bansal, S., Carlson, C., and Kraemer, J. (2020). There is no safe way to reopen colleges this fall: Reopening colleges during a pandemic is too dangerous. *Washington Post*.

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., and Deckmyn., A. (2018). *maps: Draw Geographical Maps*. R package version 3.3.0.

Britton, T., Ball, F., and Trapman, P. (2020). A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science*, 369(6505):846–849.

CDC (2021). *COVID-19*.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). *shiny: Web Application Framework for R*. R package version 1.6.0.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

D'Arienzo, M. and Coniglio, A. (2020). Assessment of the SARS-CoV-2 basic reproduction number, r0, based on the early phase of COVID-19 outbreak in italy. *Biosafety and Health*, 2(2):57–59.

Fan, G., Yang, Z., Lin, Q., Zhao, S., Yang, L., and He, D. (2020). Decreased case fatality rate of COVID-19 in the second wave: A study in 53 countries or regions. *Transboundary and Emerging Diseases*.

Gomes, M. G. M., Aguas, R., Corder, R. M., King, J. G., Langwig, K. E., Souto-Maior, C., Carneiro, J., Ferreira, M. U., and Penha-Gonçalves, C. (2020). Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold.

Gonzalez-Reiche, A. S., Hernandez, M. M., Sullivan, M. J., Ciferri, B., Alshammary, H., Obla, A., Fabre, S., Kleiner, G., Polanco, J., Khan, Z., Alburquerque, B., van de Guchte, A., Dutta, J., Francoeur, N., Melo, B. S., Oussenko, I., Deikus, G., Soto, J., Sridhar, S. H., Wang, Y.-C., Twyman, K., Kasarskis, A., Altman, D. R., Smith, M., Sebra, R., Aberg, J., Krammer, F., García-Sastre, A., Luksza, M., Patel, G., Paniz-Mondolfi, A., Gitman, M., Sordillo, E. M., Simon, V., and van Bakel, H. (2020). Introductions and early spread of SARS-CoV-2 in the new york city area. *Science*, page eabc1917.

Huynen, M. M., Martens, P., Schram, D., Weijenberg, M. P., and Kunst, A. E. (2001). The impact of heat waves and cold spells on mortality rates in the dutch population. *Environmental Health Perspectives*, 109(5):463–470.

Ke, R., Sanche, S., Romero-Severson, E., and Hengartner, N. (2020). Estimating the reproductive number r0 of SARS-CoV-2 in the united states and eight european countries and implications for vaccination.

Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D., and Lessler, J. (2020). Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. *Annals of Internal Medicine*, 173(4):262–267.

Levin, A. T., Meyerowitz-Katz, G., Owusu-Boaitey, N., Cochran, K. B., and Walsh, S. P. (2020). Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications.

Liu, Y., , Morgenstern, C., Kelly, J., Lowe, R., and Jit, M. (2021). The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC Medicine*, 19(1).

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.

Oran, D. P. and Topol, E. J. (2020). Prevalence of asymptomatic SARS-CoV-2 infection. *Annals of Internal Medicine*, 173(5):362–367.

Perrier, V., Meyer, F., and Granjon, D. (2021). *shinyWidgets: Custom Inputs Widgets for Shiny*. R package version 0.5.7.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Randolph, H. E. and Barreiro, L. B. (2020). Herd immunity: Understanding COVID-19. *Immunity*, 52(5):737–741.

Revell, L. J. (2012). phytools: An r package for phylogenetic comparative biology (and other things).

Methods in Ecology and Evolution, 3:217–223.

Rinaldi, G. and Paradisi, M. (2020). An empirical estimate of the infection fatality rate of COVID-19 from the first italian outbreak.

Roques, L., Klein, E. K., Papaïx, J., Sar, A., and Soubeyrand, S. (2020). Using early data to estimate the actual infection fatality ratio from COVID-19 in france. *Biology*, 9(5):97.

Rosenberg, E. S., Dufort, E. M., Blog, D. S., Hall, E. W., Hoefer, D., Backenson, B. P., Muse, A. T., Kirkwood, J. N., George, K. S., Holtgrave, D. R., Hutton, B. J., Zucker, H. A., Anand, M., Kaufman, A., Kuhles, D., Maxted, A., Newman, A., Pulver, W., Smith, L., Sommer, J., White, J., Dean, A., Derbyshire, V., Egan, C., Fuschino, M., Griesemer, S., Hull, R., Lamson, D., Laplante, J., McDonough, K., Mitchell, K., Musser, K., Nazarian, E., Popowich, M., Taylor, J., Walsh, A., Amler, S., Huang, A., Recchia, R., Whalen, E., Lewis, E., Friedman, C., Carrera, S., Eisenstein, L., DeSimone, A., Morne, J., Johnson, M., Navarette, K., Kumar, J., Ostrowski, S., Mazeau, A., Dreslin, S., Yates, N., Greene, D., Heslin, E., Lutterloh, E., Rosenthal, E., Barranco, M., Anand, M., Kaufman, A., Kuhles, D., Maxted, A., Newman, A., Pulver, W., Smith, L., Sommer, J., White, J., Dean, A., Derbyshire, V., Egan, C., Fuschino, M., Griesemer, S., Hull, R., Lamson, D., Laplante, J., McDonough, K., Mitchell, K., Musser, K., Nazarian, E., Popowich, M., Taylor, J., Walsh, A., Amler, S., Huang, A., Recchia, R., Whalen, E., Lewis, E., Friedman, C., Carrera, S., Eisenstein, L., DeSimone, A., Morne, J., Johnson, M., Navarette, K., Kumar, J., Ostrowski, S., Mazeau, A., Dreslin, S., Yates, N., Greene, D., Heslin, E., Lutterloh, E., Rosenthal, E., and and, M. B. (2020). COVID-19 testing, epidemic features, hospital outcomes, and household prevalence, new york state—march 2020. *Clinical Infectious Diseases*, 71(8):1953–1959.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Streeck, H., Schulte, B., Kümmerer, B. M., Richter, E., Höller, T., Fuhrmann, C., Bartok, E., Dolscheid-Pommerich, R., Berger, M., Wessendorf, L., Eschbach-Bludau, M., Kellings, A., Schwaiger, A., Coenen, M., Hoffmann, P., Stoffel-Wagner, B., Nöthen, M. M., Eis-Hübinger, A. M., Exner, M., Schmithausen, R. M., Schmid, M., and Hartmann, G. (2020). Infection fatality rate of SARS-CoV2 in a super-spreading event in germany. *Nature Communications*, 11(1).

Velavan, T. P. and Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical Medicine & International Health*, 25(3):278–280.

Vermund, S. H. and Pitzer, V. E. (2020). Asymptomatic transmission and the infection fatality risk for COVID-19: Implications for school reopening. *Clinical Infectious Diseases*.

Wong, G., , Bi, Y.-H., Wang, Q.-H., Chen, X.-W., Zhang, Z.-G., Yao, Y.-G., , , , , , and and (2020). Zoonotic origins of human coronavirus 2019 (HCoV-19 / SARS-CoV-2): why is this work important? *Zoological Research*, 41(3):213–219.

Wu, S. L., Mertens, A., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Reingold, A., Arnold, B. F., Hubbard, A., and Benjamin-Chung, J. (2020). Substantial underestimation of SARS-CoV-2 infection in the united states due to incomplete testing and imperfect test accuracy.

Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., Yu, T., Wang, Y., Pan, S., Zou, X., Yuan, S., and Shang, Y. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in wuhan, china: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine*, 8(5):475–481.