

1 *covid19.Explorer*: A web application and R 2 package to explore United States COVID-19 3 data

4 Liam J. Revell^{1,2}

5 ¹Department of Biology, University of Massachusetts Boston, Boston, Massachusetts
6 02125

7 ²Facultad de Ciencias, Universidad Católica de la Santísima Concepción, Concepción,
8 Chile

9 Corresponding author:

10 Liam J. Revell^{1,2}

11 Email address: liam.revell@umb.edu

12 ABSTRACT

13 Appearing at the end of 2019, a novel virus (later identified as SARS-CoV-2) was characterized in the city
14 of Wuhan in Hubei Province, China. As of the time of writing, the disease caused by this virus (known as
15 COVID-19) has already resulted in over 3 million deaths worldwide. SARS-CoV-2 infections and deaths,
16 however, have been highly unevenly distributed among age groups, sexes, countries, and jurisdictions
17 over the course of the pandemic. Herein, I present a tool (the *covid19.Explorer* R package and web
18 application) that has been designed to explore and analyze publicly available United States COVID-19
19 infection and death data from the 2020/21 U.S. SARS-CoV-2 pandemic. The analyses and visualizations
20 that this R package and web application facilitate can help users better comprehend the geographic
21 progress of the pandemic, the effectiveness of non-pharmaceutical interventions (such as lockdowns and
22 other measures, which have varied widely among U.S. states), and the relative risks posed by COVID-19
23 to different age groups within the U.S. population. The end result is an interactive tool that will help its
24 users develop an improved understanding of the temporal and geographic dynamics of the SARS-CoV-2
25 pandemic, accessible to lay people and scientists alike.

26 INTRODUCTION

27 In 2019, a novel infectious disease was identified in Wuhan, a city of approximately 11 million residents
28 located in the Hubei Province of central China. This infectious disease, called *Coronavirus disease 2019*,
29 or COVID-19 (Velavan and Meyer, 2020), is now known to be caused by the previously unidentified
30 *severe acute respiratory syndrome coronavirus 2* or SARS-CoV-2 (?). Following the Wuhan outbreak,
31 cases of SARS-CoV-2 infection and COVID-19 death were subsequently identified in Europe, the United
32 States, and (by the time of writing) at least 192 countries worldwide. Counting from the beginning of this
33 global pandemic, there have been nearly 3.1 million confirmed COVID-19 deaths, more than 580,000 of
34 which have occurred in the United States alone (CDC, 2021).

35 R (R Core Team, 2020) is a powerful scientific computing environment and programming language
36 that is used by statisticians, data scientists, academic researchers, and students worldwide. I have built
37 a multifunctional R package (*covid19.Explorer*) and corresponding web application ([https://covid19-
38 explorer.org](https://covid19-explorer.org)). The purpose of both is to aid scientists and lay people alike to better understand the 2020/21
39 SARS-CoV-2 pandemic in the United States. Although my focus is on U.S. COVID-19 data, readers from
40 other countries might also be interested in the project – for instance, because the seasonal dynamics of
41 infection or the age distribution of mortality has been broadly similar among different affected areas of
42 the globe.

43 This R package and website are not designed to be a substitute or replacement for the many other
44 excellent software products and web tools that have been developed over the past year (e.g., Brown et al.,
45 2020; Johns Hopkins University, 2020; Reiner et al., 2020; Gu, 2020). It nonetheless contains a number

46 of different analytical approaches and methods that distinguish it from other software and web resources.

47 For example, the *covid19.Explorer* R package is the only software that I know of that allows the user
48 to specify a custom model of the infection fatality ratio (IFR, the fraction of all SARS-CoV-2 infected
49 individuals that ultimately die of COVID-19 in a given population; Blackburn et al., 2021) through time
50 and then uses this model to reconstruct daily SARS-CoV-2 infections. Although this strategy has been
51 employed by other modelers to estimate daily SARS-CoV-2 infections throughout the pandemic (most
52 notably, perhaps, by Gu, 2020 – though other modeling groups also use confirmed daily COVID-19 deaths
53 as an important lagging indicator of new infections, e.g., Reiner et al., 2020), mine is, so far as I am aware,
54 the only software that puts this model of IFR entirely under user control.

55 Likewise, the *covid19.Explorer* R package and website includes visualization methods not available
56 in other software or web resources. For instance, the *covid19.Explorer* can create a plot of U.S. state-
57 wise daily estimated infections in aggregate that is unlike any graphical representation of United States
58 SARS-CoV-2 infection data that I have seen in other software, webpages, or media sources. Similarly, the
59 package includes an ‘iceberg graph’ showing daily observed SARS-CoV-2 infections above the waterline,
60 and estimated unobserved infections below it. I have likewise never encountered a precisely identical
61 visual representation of the U.S. COVID-19 pandemic data in other electronic resources or software.

62 Lastly, it’s perhaps important to mention one thing that the *covid19.Explorer* R package most
63 adamantly *does not* do, and that is make predictions about the future. There are numerous different
64 individual scientists and research teams that have dedicated enormous effort and resources to predicting
65 the epidemic dynamics of SARS-CoV-2 in the United States and globally (e.g., Reiner et al., 2020; Gu,
66 2020) with widely varying success (e.g., Chin et al., 2020; Ioannidis et al., 2020; James et al., 2021).
67 The *covid19.Explorer* R package and site have the more modest goal of helping users develop a better
68 understanding of what *has* happened over the course of the United States SARS-CoV-2 pandemic from its
69 beginnings to the present day.

70 1 METHODS

71 1.1 Preamble

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

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

82 The data used by the various applications of the *covid19.Explorer* are all publicly available and were
83 obtained (unless otherwise indicated) from the *United States Centers for Disease Control and Prevention*
84 *National Center for Health Statistics* (<https://www.cdc.gov/nchs/>; henceforward, the CDC) or the *United*
85 *States Census Bureau* (<https://www.census.gov/>). In particular, these data consist of: provisional U.S.
86 COVID-19 death counts by sex, age, and week from the CDC; United States confirmed COVID-19 cases
87 and deaths by state through time from the CDC; weekly counts of deaths by jurisdiction and age group
88 from the CDC, 2015-present; weekly counts of deaths by state and select causes (including COVID-19)
89 from 2014-2018, 2019-2020, and 2020-2021 from the CDC; estimated population sizes by U.S. state and
90 by age, from 2010-2019 from the Census Bureau; and, finally, the geographic center of each U.S. state (to
91 be used for mapping visualizations).

92 1.2 Types of functions in *covid19.Explorer*

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

95 The first of these (exemplified by the *shiny* web application webpage tabs denominated *U.S. COVID-19*
96 *infections*, *Iceberg plot*, *State comparison*, *Plausible range*, and *Infection estimator*) consists of functions
97 that are designed to estimate the true number of COVID-19 infections over the course of the pandemic.
98 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 are set by the user, rather than estimated from the data. In particular, 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. Each of these model parameters have been assigned 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 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 (if applicable) of all functions in *covid19.Explorer* are given in alphabetical order in Table 1, below.

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
<i>age.deaths</i>	Excess mortality by age	Graph weekly or cumulative excess mortality by age and jurisdiction.
<i>compare.infections</i>	State comparison	Compare daily or cumulative deaths and estimated daily or cumulative infections between states and U.S. jurisdictions.
<i>covid.deaths</i>	Deaths by age	Plot weekly or cumulative confirmed COVID-19 deaths by age group and compared to all deaths.
<i>iceberg.plot</i>	Iceberg plot	Graph observed daily confirmed SARS-CoV-2 cases (above the ‘waterline’ of the graph) and estimated unobserved infections (below it).
<i>infection.estimator</i>	Infection estimator	Estimate daily or cumulative SARS-CoV-2 infections based on observed deaths and confirmed cases.
<i>infection.range.estimator</i>	Plausible range	Estimate the plausible range of daily or cumulative infections based on an interval of IFR values at each time point.
<i>infections.by.state</i>	SARS-CoV-2 infections	Visualize geographic distribution of new or cumulative SARS-CoV-2 infections through time.
<i>state.deaths</i>	By state	Graph weekly or cumulative excess deaths by U.S. state.
<i>updateData</i>	Not applicable	Update the data used by <i>covid19.Explorer</i> from the web.

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 underestimation has multiple causes. One important factor is that there has been limited testing capacity throughout much of the SARS-CoV-2 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-

124 CoV-2 testing is voluntary, population surveillance testing has been relatively scarce, and many cases
125 of SARS-CoV-2 infection present asymptotically or with mild symptoms (Oran and Topol, 2020). As
126 such, I consider confirmed COVID-19 deaths to be a much more reliable indicator of disease burden than
127 confirmed cases. Deaths, however, are obviously a *lagging* indicator of infections.

128 The key parameter that relates daily COVID-19 deaths to the number of infections is the infection
129 fatality ratio (also called the *infection fatality rate* or *IFR*). IFR, normally expressed as a percent, is defined
130 as the fraction of deaths among all infected individuals, taking into account both observed infections
131 ('cases') and asymptomatic or unobserved infections (O'Driscoll et al., 2020). An IFR value of 1.5%, for
132 example, would mean that, on average, for every 1,000 infections in a specified population, there would
133 be 15 deaths.

134 I modeled the number of new SARS-CoV-2 infections on the i th day by taking the number of observed
135 COVID-19 deaths on day $i + k$ (in which k is the average lag period between initial infection and death,
136 where death is the outcome of infection), and then dividing this quantity by the IFR. In other words, given
137 50 COVID-19 deaths on day $i + k$, and an IFR of 0.5%, we would predict that 10,000 new SARS-CoV-2
138 infections *had occurred* on day i . Both k , the average lag time from infection to death (in cases of
139 SARS-CoV-2 infections resulting in death), and the IFR are to be specified by the user.

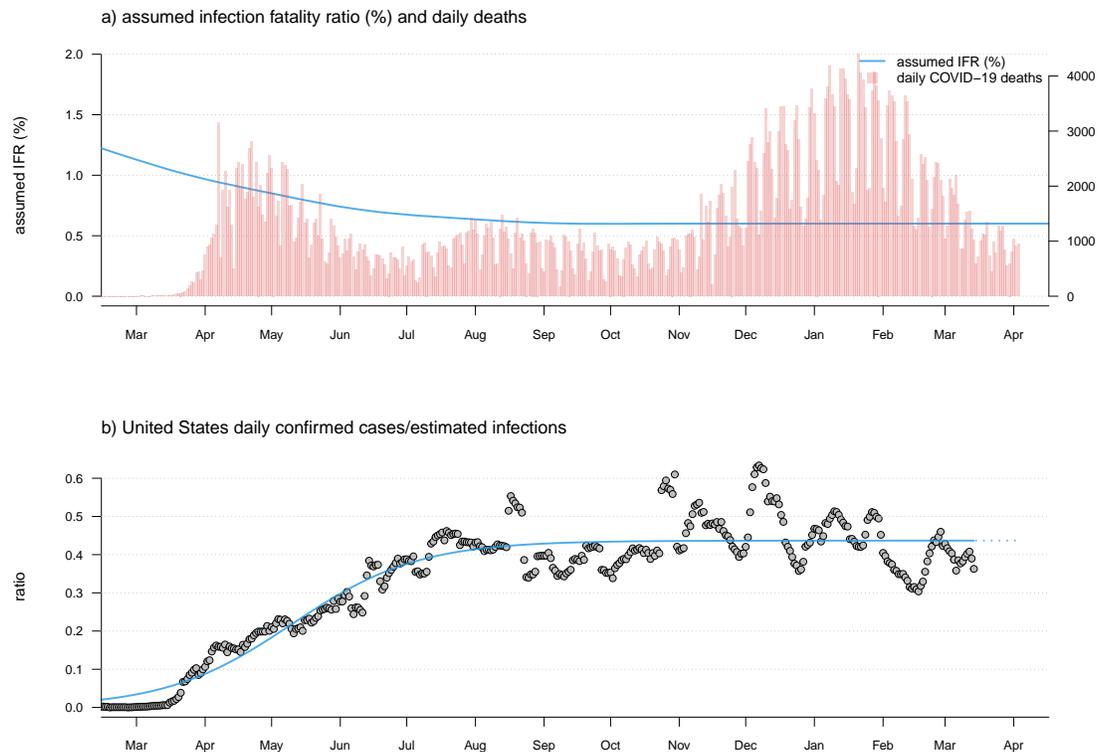


Figure 1. a) Observed U.S. daily COVID-19 deaths (red bars) and user-specified infection fatality rate (IFR) function (blue line) through time. Note that this panel of the figure has two vertical axes. The axis on the left shows IFR in %, corresponding to the user-specified IFR model indicated by the blue curved line. The axis on the right shows the number of new daily COVID-19 deaths, corresponding to the vertical red bars. b) The ratio of daily confirmed SARS-CoV-2 infections over estimated infections (grey points) and a fitted sigmoid function of the implied case detection rate (CDR) through time. This fitted curve is used to extrapolate the true number of daily new SARS-CoV-2 infections from reported cases in the most recent reporting days.

140 A fairly reasonable lag time between infection and death might be approximately three weeks. (Not to
141 be confused with the median lag time between symptom offset and death, e.g., Wilson et al., 2020.) For
142 example, during a large outbreak in Melbourne, Australia the time difference between the peak recorded

143 cases and peak confirmed COVID-19 deaths was around 17 days. Infected persons normally test negative
144 for the first few days following exposure (Kucirka et al., 2020), so this more or less corresponds with a
145 three week lag.

146 Likewise, IFR values ranging from about 0.2% to over 1.0% have been reported over the course of the
147 pandemic. For instance, a study based on an early, super-spreader event in Germany estimated an IFR
148 (corrected to the demographic distribution of the local population) of 0.36% (Streeck et al., 2020). Other
149 researchers have reported higher estimated IFR (e.g., Rinaldi and Paradisi, 2020). In a large meta-analysis
150 O'Driscoll et al. (2020) estimated IFR of SARS-CoV-2 infection across 45 different countries and obtained
151 median estimates ranging from 0.24% to 1.49%, with higher IFRs typically reported for countries with
152 older populations. In general, it is probably reasonable to suppose that IFR has fallen through time
153 as treatment of severely ill patients has improved (Fan et al., 2020). Likewise, even within the U.S.,
154 IFR is unlikely to be precisely the same at a given date in different jurisdictions, due to differences in
155 demographic structure between areas as well as other factors.

156 I suspect that it is within reason for users of *covid19.Explorer* to specify an IFR that is no greater than
157 about 1.5% and that declines gradually from the start of the pandemic towards the present, with a *current*
158 IFR that is perhaps around 0.3% - 0.5% (O'Driscoll et al., 2020; Blackburn et al., 2021). Nonetheless,
159 *covid19.Explorer* permits the user to specify a time-varying IFR by fixing the IFR at each quarter (on
160 the website), or at any arbitrary time interval (using the R package directly), and then interpolating daily
161 IFR between each period using local regression smoothing (LOESS; Cleveland, 1979). As such, it is
162 also possible to build a model for IFR through time that both falls *and* rises, perhaps as stresses on local
163 healthcare resources increase or decrease through time with rising and falling COVID-19 case numbers.

164 Reporting can vary through time including regularly over the course of the week. (For instance, fewer
165 COVID-19 deaths tend to be reported on the weekends compared to Monday through Friday; e.g., Figure
166 1a.) To take these reporting artifacts into account, I used both moving averages and local regression
167 (LOESS) smoothing. Both the window for the moving average and the LOESS smoothing parameter are
168 controlled by the user.

169 The approach of using only confirmed COVID-19 deaths – though robust – does not permit us to
170 estimate the true number of infections between k days ago and the present. To do this, I assumed a
171 sigmoidal relationship (by default) between time and the ratio of daily confirmed cases over the estimated
172 true number of infections – a quantity called the case detection rate or CDR (Figure 1b). Since the number
173 of confirmed cases cannot exceed the true number of new infections, logic dictates that the CDR should
174 have a value that falls between 0 and 1.

175 I decided on a sigmoidal relationship between the case detection rate and time because it seemed
176 reasonable to presume the ratio was very low early in the pandemic when confirming a new infection was
177 limited primarily by testing capacity, but that CDR has probably risen (in many localities) to a more or less
178 consistent value as testing capacity increased. Since getting tested is voluntary, and since many infections
179 of SARS-CoV-2 are asymptomatic or only mildly symptomatic, this ratio seems unlikely to rise to very
180 near 1.0 in the U.S. regardless of the availability of testing. Figure 1, created using *covid19.Explorer*,
181 shows daily confirmed cases / daily estimated infections (under our model) for all U.S. data over the
182 entire course of the pandemic to date (Figure 1b), given observed daily deaths (red bars) and assumed IFR
183 evolution through time (blue curved line; Figure 1a). Our plot seems to indicate a CDR of about 0.42
184 at the present; however, the reader should keep in mind that in practice this value is estimated separately for
185 each jurisdiction that is being analyzed, and as such might be lower in some states and higher in others,
186 even for a constant IFR value or function.

187 In the event that a sigmoid function cannot be fit to the implied daily CDR for a given state or
188 jurisdiction, the software automatically substitutes the mean CDR from the last 30 days of data. Since I
189 only used the CDR to estimate daily infections for the most recent time period of our data (see below), and
190 since CDR tended to increase asymptotically towards a more or less constant value in most jurisdictions
191 (e.g., Figure 1), this seemed fairly reasonable. When using the *covid19.Explorer* in R (rather than through
192 the web interface), this option can also be selected explicitly by the user. An important point to make in
193 this context is that I intend the sigmoidal functional form to be a heuristic (rather than literal) means of
194 capturing the approximate relationship between CDR and time since the start of the pandemic – and thus
195 estimate the CDR for the most recently reported cases. If users are unsatisfied with the fit of the sigmoid
196 curve to CDR, they are encouraged to substitute the mean implied CDR from the last 30 days of data. The
197 reason I chose the sigmoid fit to begin with was primarily to avoid distortions driven by so-called ‘data

198 dumps,' in which a state or jurisdiction releases a large number of previously misclassified or unreported
199 cases or deaths on a single day. In practice, using the mean implied CDR from the past 30 days or the
200 fitted value of CDR from a sigmoid fit will not make much of a difference in the majority of jurisdictions
201 represented in our data.

202 After fitting this sigmoidal curve to our observed and estimated cases through now $-k$ days (or
203 calculating the mean implied CDR from the most recent 30 days), we then must turn to the last period. To
204 obtain estimated infections for these days, we merely divide our observed cases from the last k days of
205 data by the fitted CDR values of our curve. Figure 2 shows the result of this analysis applied to data for
206 the U.S. state of Massachusetts.

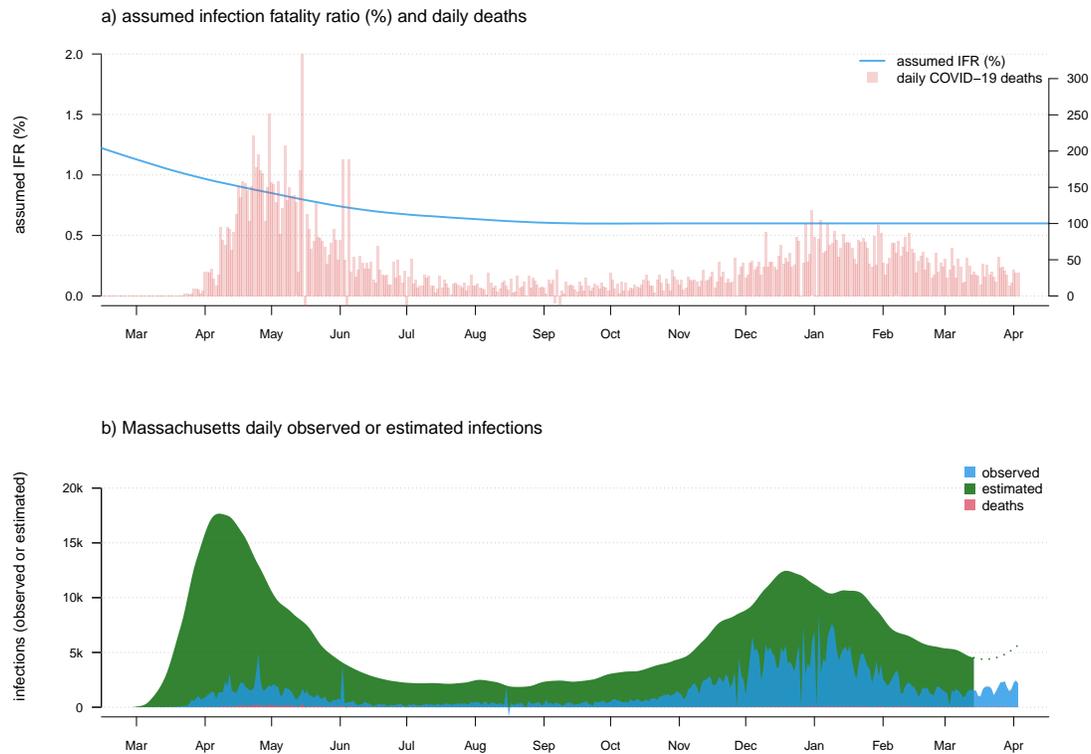


Figure 2. a) Observed daily COVID-19 deaths and an assumed model of IFR in which the infection fatality ratio is initially high (1.5%), but then declines and stabilizes at around 0.6% through the present day. As in Figure 1, panel a) has two vertical axes. The axis on the left shows IFR in %, corresponding to the user-specified IFR model indicated by the blue curved line. The axis on the right shows the number of new daily COVID-19 deaths, corresponding to the vertical red bars of the plot. b) Estimated daily infections (green), cases (blue), and deaths (red).

207 In addition to computing the raw number of daily infections, this method can also be used to estimate
208 infections as a percentage of the total population. To make this calculation, I obtained state populations
209 through time from the U.S. Census Bureau. Data was only given through 2019 at the time of writing, so
210 to estimate state-level 2020 population sizes, I used a total mid-year 2020 U.S. population estimate of
211 331,002,651 to 'correct' each 2019 state population size to a 2020 level. Finally, CDC mortality data
212 splits New York City (NYC) from the rest of New York state. Since this contrast is interesting (e.g.,
213 Gonzalez-Reiche et al., 2020), I maintained the separation – and used a mid-2019 population estimate of
214 (8,336,817) for NYC, then simply assumed that the population of NYC has changed between 2015 and
215 2020 in proportion to the rest of the state. (Since they have a part : whole relationship, this seemed pretty
216 reasonable. In fact, according to the U.S. Census Bureau from 2010 to 2019 the fraction of New York
217 State residents living in New York City is estimated to have grown by around 0.1% per year, from 41.8%
218 in 2010 to about 42.8% in 2019. If this trend continued through 2020, then I may have underestimated

219 the population of New York City by about 0.2%. Since this is only relevant when considering per capita
220 SARS-CoV-2 infections and COVID-19 deaths, I suspect it is a relatively minor source of error compared
221 to other simplifying assumptions of this software.)

222 **1.4 Assumptions about estimating infections**

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

228 Although the model does not pre-suppose a specific value or function for IFR, it does require that
229 one be specified by the user. As such, it is probably worth mentioning the effect of setting an IFR value
230 that is either too high or too low compared to the (invariably unknown) *true* IFR for the population
231 of interest. An IFR that is too high (overall or at a specific time during the pandemic) will have the
232 general effect of causing us to systematically underestimate the number of infections that have occurred.
233 This makes sense because if we imagine observing 50 COVID-19 deaths, an IFR of 0.5% would imply
234 that these deaths correspond to a total of 10,000 SARS-CoV-2 infections. By contrast, a higher IFR of,
235 say, 1.0% would instead imply that only 5,000 infections had occurred. Assuming an IFR value that is
236 too low will (obviously) have exactly the opposite effect and thus cause us to overestimate the number
237 of infections that have occurred. The default values for IFR through time specified in the web portal
238 (<https://covid19-explorer.org>) are 0.85% on Feb. 1, 2020 and then decline every 3 months: 0.65%, 0.55%,
239 0.5%, and 0.5% on January 31, 2021, with intermediate values interpolated using LOESS smoothing.

240 The purpose of the software and web resource is to allow the user to explore alternative (reasonable)
241 scenarios for IFR through time and examine their effects on estimated daily or cumulative SARS-CoV-2
242 infections in different jurisdictions; however, the default values are not arbitrary. First, they are largely
243 consistent with population-wise IFR estimates from seroprevalence research (e.g., O’Driscoll et al., 2020).
244 Second, they yield estimated daily infections that are qualitatively if not quantitatively similar to those
245 obtained by several other leading models of the SARS-CoV-2 pandemic in the United States (e.g., Gu,
246 2020; Reiner et al., 2020).

247 I also assume a homogeneous value of k at any particular time. In fact, literature sources report
248 lag-times between two and eight weeks (e.g., Yang et al., 2020). Nonetheless, I suspect that inferences
249 by this method should not be badly off – so long as the true IFR does not swing about wildly from day
250 to day, and so long as the number of deaths is not extremely few for any reporting period. I likewise
251 assume a constant lag-period, k , through time. This assumption is perhaps a bit more dubious as it seems
252 quite reasonable to suppose that, for a specific state or jurisdiction, as IFR falls k might also increase. If
253 k increased as a function of time, this would mean that recent peaks in daily new infections would be
254 systematically biased forward in time (that is, they occurred earlier than it seems) compared to peaks that
255 occurred early in the pandemic. (The converse would also be true if k decreased rather than increasing
256 through time.) This is a complexity that I explicitly chose to ignore in the model.

257 I assume that a more or less consistent fraction of COVID-19 deaths are reported as such – that is,
258 that COVID-19 is neither systematically under- or overreported as the cause of death at any point during
259 the course of the pandemic. A violation of this assumption is not quite as grave as it might seem, however,
260 because it can simply be ‘baked in’ to our model for IFR. For instance, if we think that COVID-19 deaths
261 were under-reported near the start of the pandemic (e.g., Weinberger et al., 2020), perhaps due to limited
262 testing capacity, this can be accommodated into our model for daily infections simply by specifying a
263 slightly lower IFR value for SARS-CoV-2 infection at that time (keeping in mind, of course, that the true
264 IFR has generally decreased through time; e.g., Levin et al., 2020).

265 In estimating the number of daily infections from k days ago to the present, we assume that the
266 relationship between time (since the first infections) and the ratio of confirmed and estimated infections
267 (i.e., the case detection rate, CDR) is sigmoidal in shape (Figure 1b). This is a testable assumption that
268 seems to hold fairly well across the entire U.S. (Figure 1b) and for some jurisdictions, but less well for
269 others. It is equally plausible to suspect that CDR could shift not only as a function of time, but also as
270 demands on testing capacity rise and fall with case numbers, or as different populations become infected.
271 This should be the subject of additional study, but my suspicion is that this would not be likely to have a
272 large effect on our model compared to other simplifications. Additionally (as mentioned above), when

273 using *covid19.Explorer* from within R it is straightforward to substitute the mean implied CDR from the
274 last 30 days for the fitted values of CDR from the sigmoidal fit.

275 One slightly problematic possibility is that the true CDR in the most recent k days is much lower or
276 higher than estimated CDR. This could happen if, for example, in jurisdictions with low surveillance
277 testing, changes in the demographic distribution of new SARS-CoV-2 infections (due to, for instance,
278 age-prioritized vaccination) mean that relatively few infections present symptomatically and get tested.
279 This would have the effect of causing CDR to be overestimated and would result in a concomitant
280 *underestimation* of daily new SARS-CoV-2 infections towards the right side of the graph. The opposite
281 effect is expected if surveillance testing was to be increased (for instance, in a jurisdiction with high
282 numbers of in-person college or university students simultaneously returning to campus), thus increasing
283 true CDR relative to its estimated value in the most recent period compared to time periods prior to k days
284 before the present.

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

291 **1.5 Showing observed and estimated unobserved infections using an ‘iceberg plot’**

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

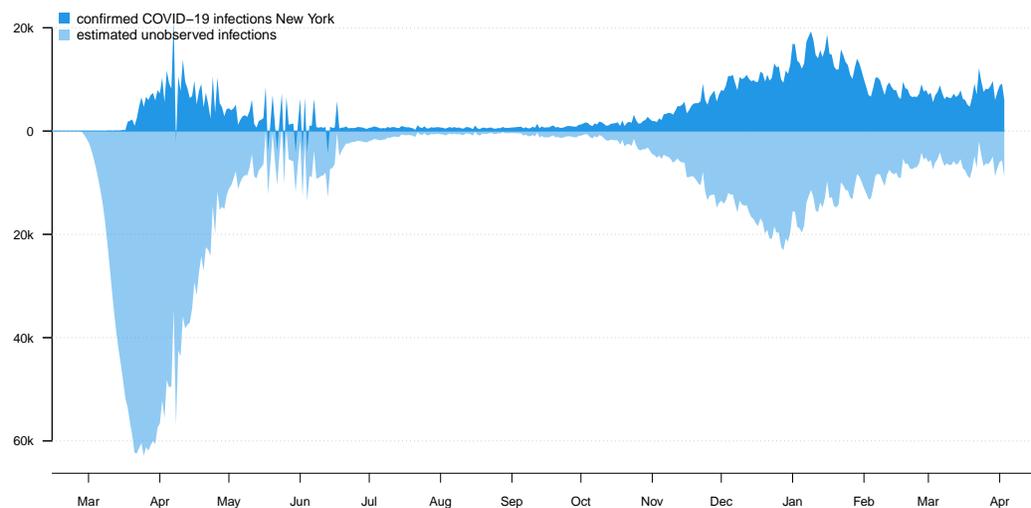


Figure 3. Iceberg plot showing the confirmed daily new infections (above the ‘waterline’ of the plot) and estimated unobserved infections (below it) for New York state.

298 **1.6 Mapping the distribution of infections across states**

299 A hallmark feature of the U.S. COVID-19 pandemic has been the shifting geographic distribution
300 of infections through time among states. To capture this dynamic, I devised a plotting method for

301 *covid19.Explorer* in which I overlay the daily or cumulative SARS-CoV-2 infections under our model
302 (outlined above), separated by state.

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

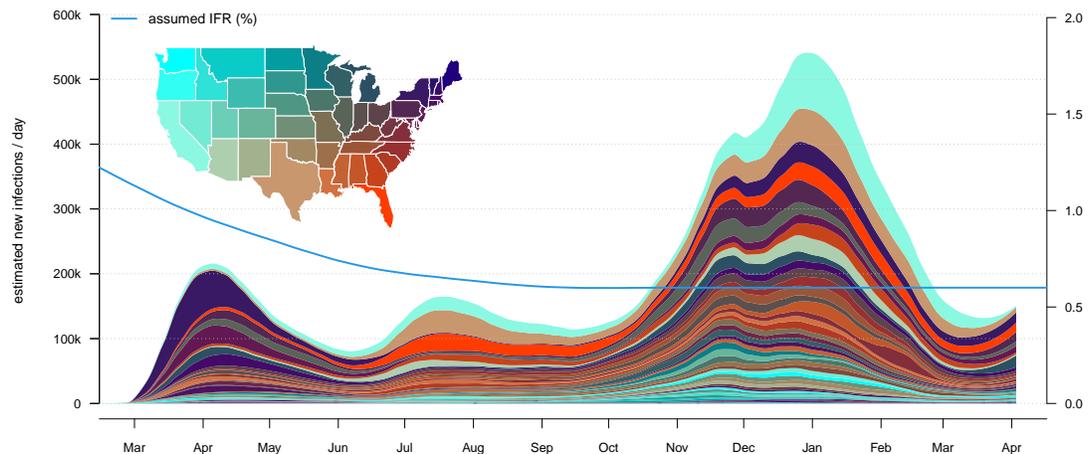


Figure 4. Daily estimated infections separated by state. The color palette is designed to capture the geographic distribution of new infections through time, rather than the severity of the pandemic in each state.

307 This plotting method shares all the assumptions of our infection estimator, above, but adds the
308 additional assumption that our model of IFR is the same for all states. This assumption is quite dubious,
309 in fact, as IFR could be expected to rise in locations where hospital resources are overtaxed by high disease
310 burden; and, conversely, fall in hospitals where staff have more experience in treating COVID-19 patients.

311 On an individual level, IFR is also very strongly influenced by age (e.g., O’Driscoll et al., 2020), as
312 well as by other risk factors such as obesity (e.g., Kompanyets et al., 2021) and socioeconomics (e.g.,
313 Lone et al., 2021). As such, even if IFR falls through time in different jurisdictions in a similar way,
314 one would nonetheless expect to observe higher IFR in states with higher median age, higher obesity, or
315 higher poverty rates, compared to younger, less obese, and higher median income states. Although I do
316 not doubt that these nuances are important in making specific, quantitative statements about the particular
317 number of infections in each state, I nonetheless believe that my method is effective at visually capturing
318 the overall geographic dynamics of the COVID-19 pandemic in the United States.

319 One point that may be worth noting about this lattermost assumption is that use of a constant IFR
320 model across all U.S. states does not, in and of itself, have the effect of distorting the *total* number of
321 estimated new infections on each day. To see this, let’s start by imagining (for instance) 3,000 infections
322 in jurisdiction A on day 1 and 30 resultant deaths k days later (IFR of 1.0%). Meanwhile, perhaps, 2,000
323 infections have occurred in jurisdiction B on day 1, but only 10 deaths k days later (IFR of 0.5%). Using
324 the global IFR of $(30 + 10)/(3,000 + 2,000) \times 100\% = 0.8\%$ gives us the same estimate of the total
325 number of new infections on day 1 ($40/0.008 = 5,000$) whether it is applied to each jurisdiction separately,
326 or to the total number of deaths taken all together. What *is* affected, however, are the proportions of new
327 infections attributed to each jurisdiction. In the constant IFR model the number of infections attributed
328 to jurisdiction A ($30/0.008 = 3,750$) would be too few; while the number of new infections attributed
329 to jurisdiction B ($10/0.008 = 1,250$) is too many. Thus the distribution of daily new infections among
330 sites, but not their grand total across jurisdictions, can be affected by an assumption that the IFR of
331 SARS-CoV-2 (and the way that it changes through time) is the same across all of the jurisdictions in our
332 dataset.

333 1.7 Visualizing COVID-19 mortality data

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

337 Excess mortality (also called *mortality displacement*; e.g., Huynen et al., 2001) is defined as the
338 number of deaths (for any period) in excess of the ‘normal’ number of deaths for the same period. To
339 compute the raw death counts for each jurisdiction, I tabulated the 2015-2018 counts with the 2019-2020
340 provisional counts. To correct observed deaths in prior years to 2020 levels, I simply multiplied the
341 past-year death tally by the ratio the jurisdiction population in 2020 compared to the population in the
342 past year. Finally, to compute excess deaths for any jurisdiction, I then took the death counts (or corrected
343 death counts) for 2020, and subtracted the mean of years 2015 through 2019. This treats 2015 through
344 2019 as ‘normal’ years, and 2020 as unusual.

345 One factor that I did not account for in this lattermost calculation is movement of people between
346 jurisdictions. In fact, some studies indicate that the COVID-19 pandemic has disrupted normal immigration
347 patterns of humans (e.g., Smith and Wesselbaum, 2020). Areas harder-hit by SARS-CoV-2 may have
348 experienced a net loss of residents (even apart from direct mortality due to COVID-19) due to emigration
349 of people from the affected jurisdiction, or reduced immigration to the area (Smith and Wesselbaum,
350 2020). Fortunately, the *covid19.Explorer* R package and web application will be easy to update when
351 final census and estimated population sizes for the states and jurisdictions of my dataset are published for
352 2020 and 2021.

353 1.8 The *covid19.Explorer* web interface

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

358 Figure 5 shows a screenshot of this web application, illustrating an analysis of the estimated cumulative
359 number of SARS-CoV-2 infections through time across U.S. states. In this web application, the user
360 must specify the value of IFR at the beginning of each three month period, and that at the end of the year,
361 beginning on Feb. 1, 2020, and ending on Jan. 31, 2021. Values on these intervals are interpolated using
362 LOESS smoothing.

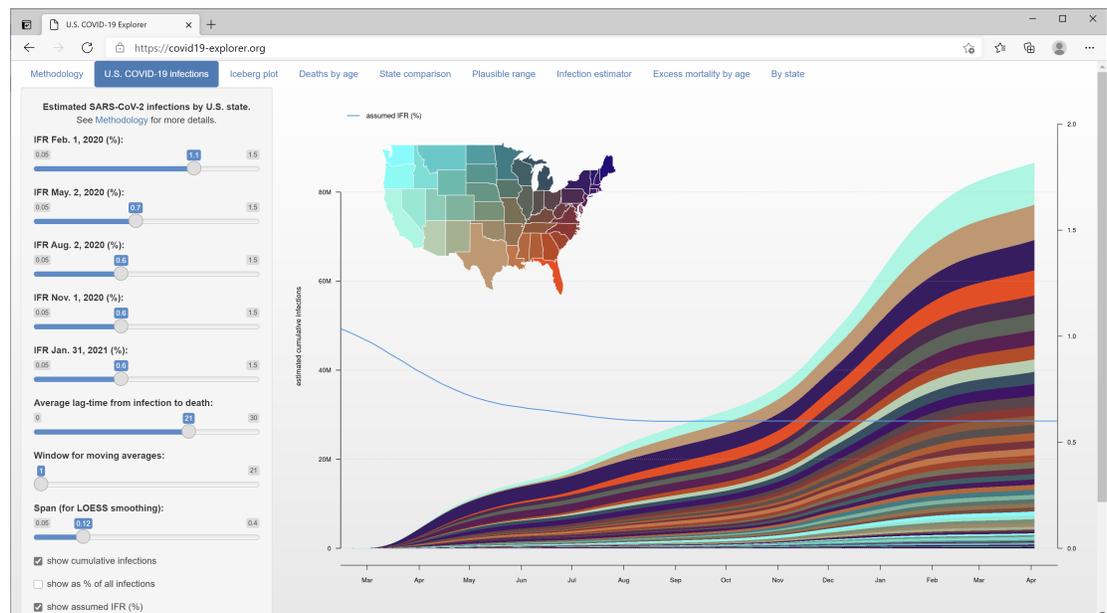


Figure 5. The *covid19.Explorer* web interface (<https://covid19-explorer.org>) showing estimated cumulative SARS-CoV-2 infections among states under the same IFR model as Figures 1 – 4.

363 Although the default values for the IFR of SARS-CoV-2 and the average lag time from infection to

364 death on the web interface are somewhat arbitrary (and are meant to be adjusted by the user), they both
365 fall on the range of most estimated values for these parameters from other research (e.g., O’Driscoll et al.,
366 2020; Wilson et al., 2020), and result in estimated daily new SARS-CoV-2 infections that are qualitatively
367 and/or quantitatively similar to other leading resources (e.g., Gu, 2020; Reiner et al., 2020).

368 2 RESULTS

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

372 2.1 Herd immunity and the cumulative proportion of the population infected

373 The question of cumulative percent infected is relevant to the (unnecessarily controversial) concept of
374 ‘herd immunity’ (Randolph and Barreiro, 2020). The herd immunity threshold (HIT), whether reached
375 via natural infection or vaccination, is typically defined as the proportion of the population that must
376 be immune in order to cause the basic reproductive number of the virus at time t (R_t) to fall below 1.0,
377 absent mitigations (Anderson and May, 1985). When R_t has fallen below 1.0, daily new infections should
378 progressively decline.

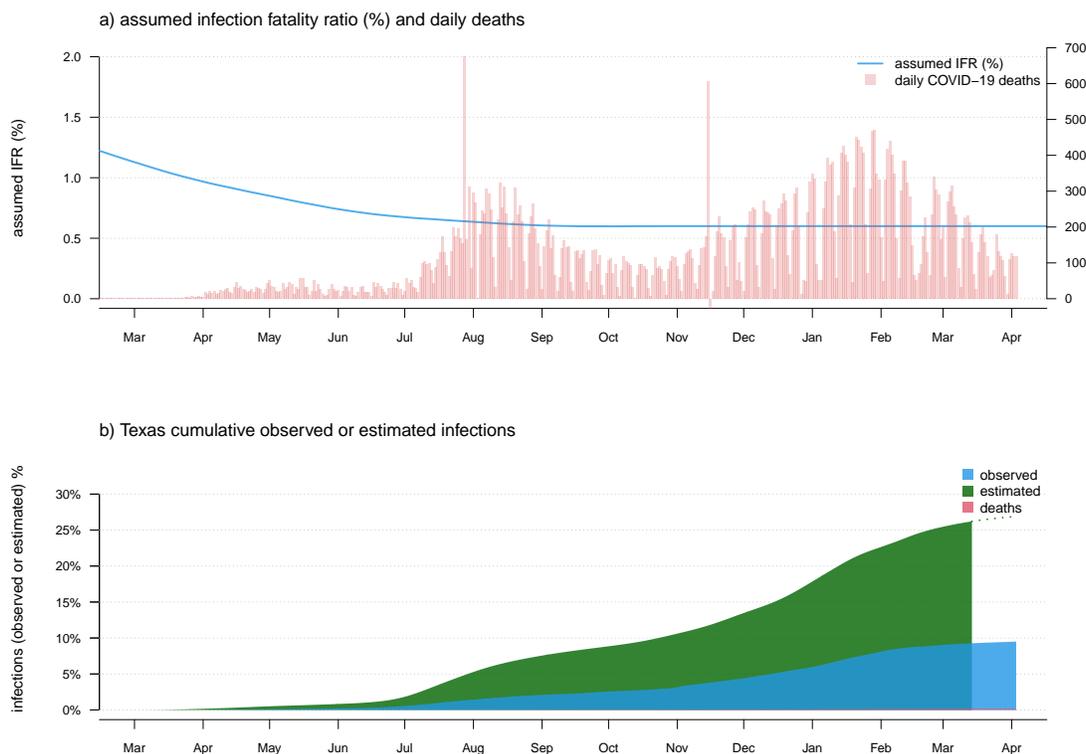


Figure 6. a) Observed daily COVID-19 deaths and an assumed model of IFR. As in Figure 1 and 2, panel a) has two vertical axes. The axis on the left shows IFR in %, corresponding to the user-specified IFR model indicated by the blue curved line. The axis on the right shows the number of new daily COVID-19 deaths, corresponding to the vertical red bars of the plot. b) Estimated cumulative SARS-CoV-2 infections (green), cases (blue), and deaths (red), as a percentage of the total population of the state.

379 The HIT is normally estimated by taking the reproductive number when 100% of the population is
380 susceptible (i.e., when a new disease emerges, R_0), and computing $1 - 1/R_0$. For SARS-CoV-2 various
381 values of R_0 have been represented in the literature, from as low as around $R_0 = 2.4$ (e.g., D’Arienzo

382 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,
383 for example, would imply that herd immunity should be reached after $1 - 1/3$ or around 67% of the
384 population has acquired immunity through natural infection or vaccination (not accounting for waning
385 acquired immunity from natural infection, which some studies have indicated for SARS-CoV-2; e.g.,
386 Long et al., 2020).

387 The *covid19.Explorer* R package and web application can be used to evaluate the proportion of
388 individuals in the total population that have been potentially infected with SARS-CoV-2, given our model
389 for COVID-19 IFR through time. Figure 6 shows cumulative estimated SARS-CoV-2 infections as a
390 fraction of the total population for the U.S. state of Texas, using the same IFR model as in Figures 1, 2,
391 3, and 4. Though the plot suggests that perhaps around 25-30% of the population in Texas has already
392 been infected, users should keep in mind that this result is entirely dependent on how we decided to
393 specify our model of IFR through time! Likewise, though this fraction is considerable, it is still well
394 below the level of infection (e.g., 67%) required to achieve herd immunity given the majority of published
395 estimates for R_0 of SARS-CoV-2. It may be worth noting that some authors have pointed out that the herd
396 immunity threshold from a natural epidemic could be considerably lower than the $1 - 1/R_0$ level expected
397 for random vaccination (e.g., Britton et al., 2020; Gomes et al., 2020). This is an intriguing possibility,
398 and one that could be qualitatively examined with some of the tools of the *covid19.Explorer* package.

399 **2.2 Computing a plausible range of infection numbers**

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

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

408 It should be noted that although the shaded region around the mean number of daily or cumulative
409 infections in Figure 7 looks like a confidence band, it would only be valid to consider it as such if our
410 high and low values of the IFR through time represented a *confidence interval* around the true infection
411 fatality rate (and, even then, this confidence band would only take into account one source of uncertainty
412 about the real daily number of infections – the IFR). As an increasing number of studies are able to
413 provide us with better and better estimates of the IFR of SARS-CoV-2 throughout this pandemic (e.g.,
414 O’Driscoll et al., 2020) it may be possible to parameterize this model in a way that genuinely accounts for
415 changing uncertainty in the value of IFR through time in the U.S. pandemic. For the time being, however,
416 I recommend employing the method as a heuristic approach to obtaining a credible range of daily new or
417 cumulative SARS-CoV-2 infections under an explicit model for the United States or any particular U.S.
418 jurisdiction.

419 **2.3 Comparing daily and cumulative infections between states**

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

429 This plotting method obviously shares *all* of the assumptions of our infection estimator, and (just like
430 our method for visualizing the geographic dynamics of the pandemic across all U.S. states) requires that
431 we use the same IFR model for each state. Since the daily and cumulative number of infections scales
432 with population size, valid state-to-state comparisons really only make sense if done on a per-capita basis
433 (e.g., infections or deaths / 1M population), just as shown here in Figure 8.

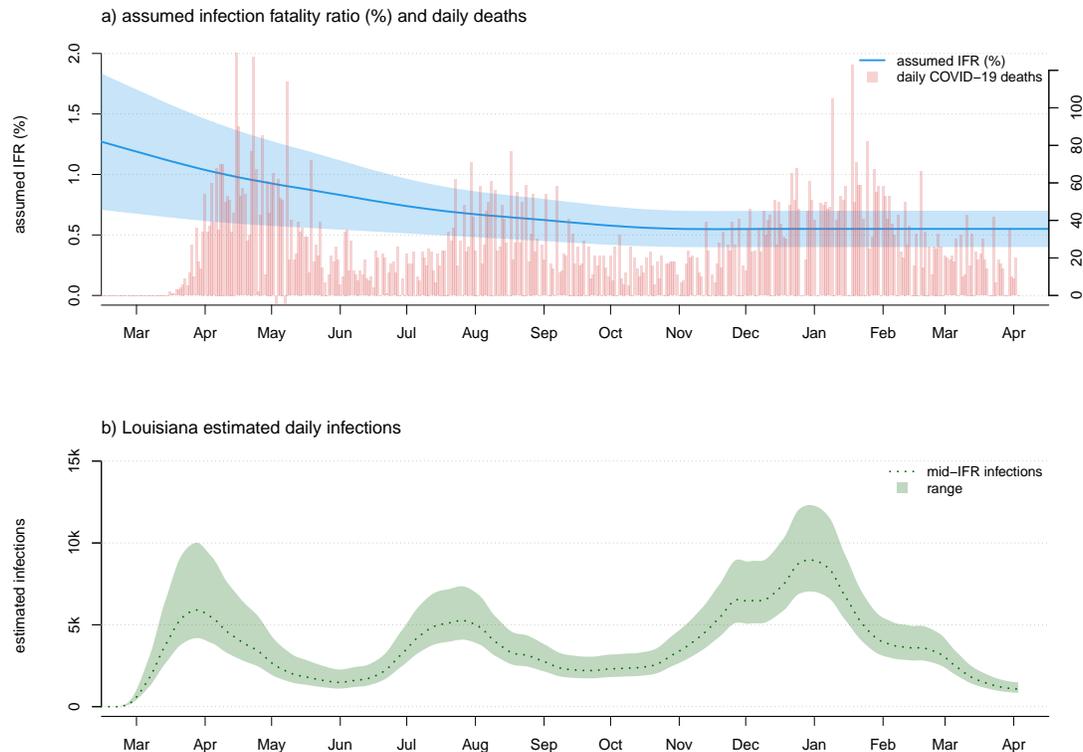


Figure 7. a) Confirmed COVID-19 deaths and a plausible range of scenarios for the evolution of SARS-CoV-2 infection fatality rate (IFR) through time. As in Figures 1, 2, and 6 panel a) has two vertical axes. The axis on the left shows IFR in %, corresponding to the user-specified IFR model (in this case, given as a plausible range of IFR values for each time period) indicated by the blue curved line and shaded area. The axis on the right shows the number of new daily COVID-19 deaths, corresponding to the vertical red bars of the plot. b) A corresponding plausible range of daily new infections, under our model, for the U.S. state of Louisiana.

434 2.4 COVID-19 mortality and age

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

438 This functionality, too, can sometimes lead to valuable insights. For instance, it was widely predicted
439 by media and public health experts that school and college reopening in the fall was likely to increased
440 SARS-CoV-2 infections and increased COVID-19 deaths among U.S. children and young people, as well
441 as increased SARS-CoV-2 transmission in the community (e.g., Bansal et al., 2020). In my opinion, the
442 minimum standard of evidence required to establish that reopening of colleges and universities for the
443 fall semester of 2020 had led to increased community transmission overall (remembering the adolescents
444 and young adults live in communities, regardless of whether they are on campus or at home) would be
445 increased SARS-CoV-2 infections of college-aged youth, as a proportion of all infections, during the fall
446 than in spring or summer.

447 In fact, and keeping in mind that COVID-19 deaths are always a better (though lagging) indicator of
448 SARS-CoV-2 infections than observed cases, CDC mortality data show precisely the opposite pattern.
449 Figure 9 gives the weekly COVID-19 deaths over all ages (in panel a) and for 15-24 year olds (in panel b).
450 We see that although the highest peaks of weekly COVID-19 deaths in the general population occurred in
451 the spring of 2020 and the fall/winter of 2020/21, peak deaths among 15-24 are similar between summer
452 and fall, and much higher (as a proportion of all COVID-19 deaths) during the summer – precisely
453 when schools and colleges were out of session for all students. This implies in turn that adolescents and

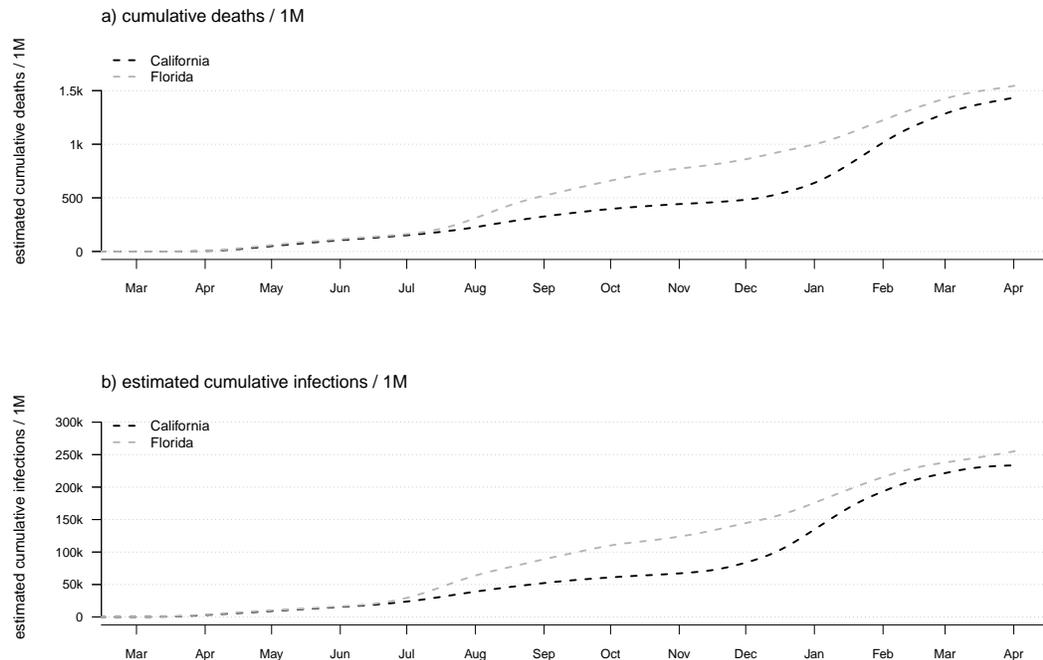


Figure 8. Daily confirmed COVID-19 deaths (a) or estimated SARS-CoV-2 infections (b) in the U.S. states of California vs. Florida.

454 college-aged adults may have been *more* (rather than less) likely to become infected with SARS-CoV-2
455 when schools were in summer recess then when they returned to campus in the fall.

456 3 DISCUSSION

457 The SARS-CoV-2 global pandemic of 2020 and 2021 has upended economies and civil society worldwide.
458 With widespread vaccination campaigns underway in many countries, and particularly in the United
459 States, the COVID-19 pandemic may *finally* be in its waning days (even if SARS-CoV-2 ultimate becomes
460 endemic and never entirely goes away, e.g., Shaman and Galanti, 2020). Nonetheless, understanding
461 the temporal and geographical dynamics of SARS-CoV-2 infections and COVID-19 deaths remains a
462 critically important endeavor. The COVID-19 pandemic is neither the first, nor will it be the last, global
463 respiratory virus pandemic (Saunders-Hastings and Krewski, 2016; Piret and Boivin, 2021). Lessons
464 learned from this pandemic will be of substantial and lasting consequence in managing or failing to
465 manage future public health emergencies.

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

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

477 Firstly, no other software or web application, to my knowledge, lets the user build a *custom model* for
478 the evolution of infection fatality rate through time. This facility, offered by *covid19.Explorer*, allows
479 the scientists and lay people that interact with the software to design their own parameter function (be it

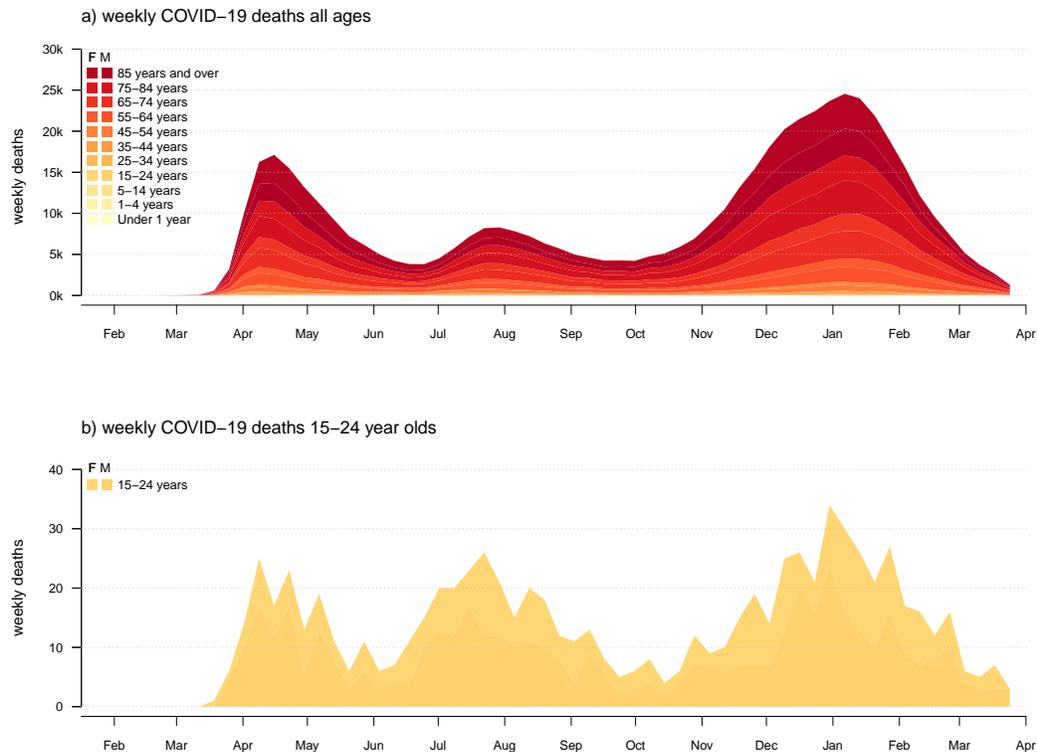


Figure 9. Weekly confirmed COVID-19 deaths for (a) all ages; and (b) individuals aged 15-24 years old in the U.S.

480 based on specific hypothesis about IFR through time, or external information – e.g., from seroprevalence
481 studies – about the value of IFR for SARS-CoV-2 at a specific time and place) that will then be used to
482 estimate infections under the model. Likewise, the tool allows *covid19.Explorer* users to progressively
483 adjust the parameter values and other assumptions of this model and see how their results change in turn.

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

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

494 Even if the SARS-CoV-2 pandemic eventually becomes a distant memory, I hope that this tool (which
495 I plan to make available indefinitely) will continue to be of use to scientists and educated lay people
496 interested in the learning from the successes and failures of policy during the 2020/21 pandemic – perhaps
497 to ensure that there are more of the former and fewer of the latter in our next global infectious disease
498 pandemic.

499 ACKNOWLEDGMENTS

500 This work is inspired by the brilliant, independent COVID-19 research of Y. Gu; by the calm and
501 thoughtful public health insights of J. Allen, F. Balloux, S. Baral, M. Gandhi, A. Munro, and others like
502 them; by my wife, E. Lu, who has been forced (due to stay at home orders and other restrictions) to suffer

503 much more of my research struggles than she would under normal circumstances; and by the persistent
504 failure of governments, public officials, and private citizens worldwide to make evidence-based decisions
505 that take into account real risks and collateral harms, in favor of unscientific performative public health
506 theater designed to abate fear. The article was improved greatly over earlier versions due to numerous
507 helpful comments from N. Dimonaco, A. Kala, and one anonymous reviewer.

508 REFERENCES

- 509 Al-Sadeq, D. W. and Nasrallah, G. K. (2020). The incidence of the novel coronavirus SARS-CoV-2
510 among asymptomatic patients: A systematic review. *International Journal of Infectious Diseases*,
511 98:372–380.
- 512 Ammar, R. (2019). *randomcoloR: Generate Attractive Random Colors*. R package version 1.1.0.1.
- 513 Anderson, R. M. and May, R. M. (1985). Vaccination and herd immunity to infectious diseases. *Nature*,
514 318(6044):323–329.
- 515 Bansal, S., Carlson, C., and Kraemer, J. (2020). There is no safe way to reopen colleges this fall:
516 Reopening colleges during a pandemic is too dangerous. *Washington Post*.
- 517 Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., and Deckmyn, A. (2018). *maps: Draw*
518 *Geographical Maps*. R package version 3.3.0.
- 519 Bennett, M. (2021). All things equal? Heterogeneity in policy effectiveness against COVID-19 spread in
520 Chile. *World Development*, 137:105208.
- 521 Blackburn, J., Yiannoutsos, C. T., Carroll, A. E., Halverson, P. K., and Menachemi, N. (2021). Infection
522 fatality ratios for COVID-19 among noninstitutionalized persons 12 and older: Results of a random-
523 sample prevalence study. *Annals of Internal Medicine*, 174(1):135–136.
- 524 Britton, T., Ball, F., and Trapman, P. (2020). A mathematical model reveals the influence of population
525 heterogeneity on herd immunity to SARS-CoV-2. *Science*, 369(6505):846–849.
- 526 Brown, M., Curiskis, A., French, A., Glickhouse, R., Goldfarb, A., Kodysh, J., Lipton, Z., Luo,
527 D., Malaty-Rivera, J., Mart, M., and et al. (2020). The COVID tracking project by The Atlantic.
528 <https://covidtracking.com/>.
- 529 CDC (2021). *COVID-19*.
- 530 Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A.,
531 and Borges, B. (2021). *shiny: Web Application Framework for R*. R package version 1.6.0.
- 532 Chin, V., Samia, N. I., Marchant, R., Rosen, O., Ioannidis, J. P. A., Tanner, M. A., and Cripps, S. (2020).
533 A case study in model failure? COVID-19 daily deaths and ICU bed utilisation predictions in New
534 York state. *European Journal of Epidemiology*, 35(8):733–742.
- 535 Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the*
536 *American Statistical Association*, 74(368):829–836.
- 537 D'Arienzo, M. and Coniglio, A. (2020). Assessment of the SARS-CoV-2 basic reproduction number, R_0 ,
538 based on the early phase of COVID-19 outbreak in Italy. *Biosafety and Health*, 2(2):57–59.
- 539 Fan, G., Yang, Z., Lin, Q., Zhao, S., Yang, L., and He, D. (2020). Decreased case fatality rate of
540 COVID-19 in the second wave: A study in 53 countries or regions. *Transboundary and Emerging*
541 *Diseases*.
- 542 Gomes, M. G. M., Aguas, R., Corder, R. M., King, J. G., Langwig, K. E., Souto-Maior, C., Carneiro, J.,
543 Ferreira, M. U., and Penha-Gonçalves, C. (2020). Individual variation in susceptibility or exposure to
544 SARS-CoV-2 lowers the herd immunity threshold.
- 545 Gonzalez-Reiche, A. S., Hernandez, M. M., Sullivan, M. J., Ciferri, B., Alshammery, H., Obla, A., Fabre,
546 S., Kleiner, G., Polanco, J., Khan, Z., Albuquerque, B., van de Guchte, A., Dutta, J., Francoeur, N.,
547 Melo, B. S., Oussenko, I., Deikus, G., Soto, J., Sridhar, S. H., Wang, Y.-C., Twyman, K., Kasarskis, A.,
548 Altman, D. R., Smith, M., Sebra, R., Aberg, J., Krammer, F., García-Sastre, A., Luksza, M., Patel, G.,
549 Paniz-Mondolfi, A., Gitman, M., Sordillo, E. M., Simon, V., and van Bakel, H. (2020). Introductions
550 and early spread of SARS-CoV-2 in the New York City area. *Science*, 369:297–301.
- 551 Gu, Y. (2020). COVID-19 projections using machine learning. <https://covid19-projections.com>.
- 552 Huynen, M. M., Martens, P., Schram, D., Weijenberg, M. P., and Kunst, A. E. (2001). The impact of heat
553 waves and cold spells on mortality rates in the Dutch population. *Environmental Health Perspectives*,
554 109(5):463–470.
- 555 Ioannidis, J. P., Cripps, S., and Tanner, M. A. (2020). Forecasting for COVID-19 has failed. *International*
556 *Journal of Forecasting*.

- 557 James, L. P., Salomon, J. A., Buckee, C. O., and Menzies, N. A. (2021). The use and misuse of
558 mathematical modeling for infectious disease policymaking: Lessons for the COVID-19 pandemic.
559 *Medical Decision Making*, page 0272989X2199039.
- 560 Johns Hopkins University (2020). Johns Hopkins University Center for Systems Science and Engineering
561 COVID-19 Dashboard.
- 562 Ke, R., Sanche, S., Romero-Severson, E., and Hengartner, N. (2020). Estimating the reproductive number
563 R_0 of SARS-CoV-2 in the United States and eight European countries and implications for vaccination.
- 564 Kompaniyets, L., Goodman, A. B., Belay, B., Freedman, D. S., Sucusky, M. S., Lange, S. J., Gundlapalli,
565 A. V., Boehmer, T. K., and Blanck, H. M. (2021). Body Mass Index and Risk for COVID-19–related
566 hospitalization, intensive care unit admission, invasive mechanical ventilation, and death — United
567 States, March–December 2020. *MMWR. Morbidity and Mortality Weekly Report*, 70(10):355–361.
- 568 Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D., and Lessler, J. (2020). Variation in false-
569 negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since
570 exposure. *Annals of Internal Medicine*, 173(4):262–267.
- 571 Levin, A. T., Meyerowitz-Katz, G., Owusu-Boaitey, N., Cochran, K. B., and Walsh, S. P. (2020). Assessing
572 the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and
573 public policy implications.
- 574 Liu, Y., , Morgenstern, C., Kelly, J., Lowe, R., and Jit, M. (2021). The impact of non-pharmaceutical
575 interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC Medicine*, 19(1).
- 576 Lone, N. I., McPeake, J., Stewart, N. I., Blayney, M. C., Seem, R. C., Donaldson, L., Glass, E., Haddow,
577 C., Hall, R., Martin, C., Paton, M., Smith-Palmer, A., Kaye, C. T., and Puxty, K. (2021). Influence of
578 socioeconomic deprivation on interventions and outcomes for patients admitted with COVID-19 to
579 critical care units in Scotland: A national cohort study. *The Lancet Regional Health - Europe*, 1:100005.
- 580 Long, Q.-X., Tang, X.-J., Shi, Q.-L., Li, Q., Deng, H.-J., Yuan, J., Hu, J.-L., Xu, W., Zhang, Y., Lv, F.-J.,
581 Su, K., Zhang, F., Gong, J., Wu, B., Liu, X.-M., Li, J.-J., Qiu, J.-F., Chen, J., and Huang, A.-L. (2020).
582 Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nature Medicine*,
583 26(8):1200–1204.
- 584 Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.
- 585 O’Driscoll, M., Santos, G. R. D., Wang, L., Cummings, D. A. T., Azman, A. S., Paireau, J., Fontanet, A.,
586 Cauchemez, S., and Salje, H. (2020). Age-specific mortality and immunity patterns of SARS-CoV-2.
587 *Nature*, 590(7844):140–145.
- 588 Oran, D. P. and Topol, E. J. (2020). Prevalence of asymptomatic SARS-CoV-2 infection. *Annals of*
589 *Internal Medicine*, 173(5):362–367.
- 590 Perrier, V., Meyer, F., and Granjon, D. (2021). *shinyWidgets: Custom Inputs Widgets for Shiny*. R package
591 version 0.5.7.
- 592 Piret, J. and Boivin, G. (2021). Pandemics throughout history. *Frontiers in Microbiology*, 11.
- 593 R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for
594 Statistical Computing, Vienna, Austria.
- 595 Randolph, H. E. and Barreiro, L. B. (2020). Herd immunity: Understanding COVID-19. *Immunity*,
596 52(5):737–741.
- 597 Reiner, B., Barber, R., and Murray, C. J. L. (2020). Modeling COVID-19 scenarios for the united states.
598 *Nature Medicine*, 27(1):94–105.
- 599 Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things).
600 *Methods in Ecology and Evolution*, 3:217–223.
- 601 Rinaldi, G. and Paradisi, M. (2020). An empirical estimate of the infection fatality rate of COVID-19
602 from the first italian outbreak.
- 603 Roques, L., Klein, E. K., Papaix, J., Sar, A., and Soubeyrand, S. (2020). Using early data to estimate the
604 actual infection fatality ratio from COVID-19 in france. *Biology*, 9(5):97.
- 605 Rosenberg, E. S., Dufort, E. M., Blog, D. S., Hall, E. W., Hoefler, D., Backenson, B. P., Muse, A. T.,
606 Kirkwood, J. N., George, K. S., Holtgrave, D. R., Hutton, B. J., Zucker, H. A., Anand, M., Kaufman,
607 A., Kuhles, D., Maxted, A., Newman, A., Pulver, W., Smith, L., Sommer, J., White, J., Dean, A.,
608 Derbyshire, V., Egan, C., Fuschino, M., Griesemer, S., Hull, R., Lamson, D., Laplante, J., McDonough,
609 K., Mitchell, K., Musser, K., Nazarian, E., Popowich, M., Taylor, J., Walsh, A., Amler, S., Huang, A.,
610 Recchia, R., Whalen, E., Lewis, E., Friedman, C., Carrera, S., Eisenstein, L., DeSimone, A., Morne, J.,
611 Johnson, M., Navarette, K., Kumar, J., Ostrowski, S., Mazeau, A., Dreslin, S., Yates, N., Greene, D.,

- 612 Heslin, E., Lutterloh, E., Rosenthal, E., Barranco, M., Anand, M., Kaufman, A., Kuhles, D., Macted,
613 A., Newman, A., Pulver, W., Smith, L., Sommer, J., White, J., Dean, A., Derbyshire, V., Egan, C.,
614 Fuschino, M., Griesemer, S., Hull, R., Lamson, D., Laplante, J., McDonough, K., Mitchell, K., Musser,
615 K., Nazarian, E., Popowich, M., Taylor, J., Walsh, A., Amler, S., Huang, A., Recchia, R., Whalen, E.,
616 Lewis, E., Friedman, C., Carrera, S., Eisenstein, L., DeSimone, A., Morne, J., Johnson, M., Navarette,
617 K., Kumar, J., Ostrowski, S., Mazeau, A., Dreslin, S., Yates, N., Greene, D., Heslin, E., Lutterloh, E.,
618 Rosenthal, E., and and, M. B. (2020). COVID-19 testing, epidemic features, hospital outcomes, and
619 household prevalence, New York state—March 2020. *Clinical Infectious Diseases*, 71(8):1953–1959.
- 620 RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.
- 621 Saunders-Hastings, P. and Krewski, D. (2016). Reviewing the history of pandemic influenza: Understand-
622 ing patterns of emergence and transmission. *Pathogens*, 5(4):66.
- 623 Shaman, J. and Galanti, M. (2020). Will SARS-CoV-2 become endemic? *Science*, 370(6516):527–529.
- 624 Smith, M. D. and Wesselbaum, D. (2020). COVID-19, food insecurity, and migration. *The Journal of*
625 *Nutrition*, 150(11):2855–2858.
- 626 Streeck, H., Schulte, B., Kümmerer, B. M., Richter, E., Höller, T., Fuhrmann, C., Bartok, E., Dolscheid-
627 Pommerich, R., Berger, M., Wessendorf, L., Eschbach-Bludau, M., Kellings, A., Schwaiger, A.,
628 Coenen, M., Hoffmann, P., Stoffel-Wagner, B., Nöthen, M. M., Eis-Hübinger, A. M., Exner, M.,
629 Schmithausen, R. M., Schmid, M., and Hartmann, G. (2020). Infection fatality rate of SARS-CoV2 in
630 a super-spreading event in germany. *Nature Communications*, 11(1).
- 631 Velavan, T. P. and Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical Medicine & International*
632 *Health*, 25(3):278–280.
- 633 Vermund, S. H. and Pitzer, V. E. (2020). Asymptomatic transmission and the infection fatality risk for
634 COVID-19: Implications for school reopening. *Clinical Infectious Diseases*.
- 635 Weinberger, D. M., Chen, J., Cohen, T., Crawford, F. W., Mostashari, F., Olson, D., Pitzer, V. E., Reich,
636 N. G., Russi, M., Simonsen, L., Watkins, A., and Viboud, C. (2020). Estimation of excess deaths
637 associated with the COVID-19 pandemic in the united states, march to may 2020. *JAMA Internal*
638 *Medicine*.
- 639 Wilson, N., Kvalsvig, A., Barnard, L. T., and Baker, M. G. (2020). Case-fatality risk estimates for
640 COVID-19 calculated by using a lag time for fatality. *Emerging Infectious Diseases*, 26(6).
- 641 Wu, S. L., Mertens, A., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S.,
642 Colford, J. M., Reingold, A., Arnold, B. F., Hubbard, A., and Benjamin-Chung, J. (2020). Substantial
643 underestimation of SARS-CoV-2 infection in the united states due to incomplete testing and imperfect
644 test accuracy.
- 645 Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., Yu, T., Wang, Y.,
646 Pan, S., Zou, X., Yuan, S., and Shang, Y. (2020). Clinical course and outcomes of critically ill patients
647 with SARS-CoV-2 pneumonia in wuhan, china: a single-centered, retrospective, observational study.
648 *The Lancet Respiratory Medicine*, 8(5):475–481.