

1 **Herd immunity vs suppressed equilibrium in COVID-19 pandemic: different**
2 **goals require different models for tracking**

3

4

5 **Authors:**

6 Norden E. Huang[#], Fangli Qiao[#], Wang Qian⁺ and Ka-Kit Tung^{*}

7

8 **Affiliations:**

9 [#] *Data Analysis Laboratory, FIO, Qingdao 266061, China*

10 ⁺ *Shanghai Jiao Tong University, Shanghai, 200240, China*

11 ^{*} *Department of Applied Mathematics, University of Washington, Seattle, WA 98195*

12 [#] Co-first authors: These authors contribute equally to this work.

13 FQ: qiaofl@fio.org.cn. NEH: norden@ncu.edu.tw

14 ^{*}Corresponding author: ktung@uw.edu

15

16 **KEYWORDS**

17 Covid-19; Epidemiology; herd immunity; suppressed equilibrium; Data-driven
18 approach; prediction of turning point; Active Infected Cases; peak total
19 hospitalizations; duration of hospital stay; end of epidemic; local-in-time metric.

20

21

22 **ABSTRACT**

23

24 **New COVID-19 epicenters have sprung up in Europe and US as the epidemic in**
25 **China wanes. Many mechanistic models' past predictions for China were**
26 **widely off the mark (1, 2), and still vary widely for the new epicenters, due to**
27 **uncertain disease characteristics. The epidemic ended in Wuhan, and later in**
28 **South Korea, with less than 1% of their population infected, much less than**
29 **that required to achieve "herd immunity". Now as most countries pursue the**
30 **goal of "suppressed equilibrium", the traditional concept of "herd immunity"**
31 **in epidemiology needs to be re-examined. Traditional model predictions of**
32 **large potential impacts serve their purpose in prompting policy decisions on**
33 **contact suppression and lockdown to combat the spread, and are useful for**
34 **evaluating various scenarios. After imposition of these measures it is**
35 **important to turn to statistical models that incorporate real-time information**
36 **that reflects ongoing policy implementation and degrees of compliance to**
37 **more realistically track and project the epidemic's course. Here we apply such**
38 **a tool, supported by theory and validated by past data as accurate, to US and**
39 **Europe. Most countries started with a Reproduction Number of 4 and declined**
40 **to around 1 at a rate highly dependent on contact-reduction measures.**

41

42

43

44

45 **1. Introduction.**

46

47 Almost one hundred years ago, a classic paper published in Proceedings of Royal
48 Society by Kermack and McKendrick (3), entitled “A contribution of mathematical
49 theory of epidemic”, started the tradition of mathematical modelling on the spread
50 of infectious disease among a susceptible population. Current thinking in
51 epidemiology is still deeply rooted in concepts introduced in that paper, some of
52 which are still relevant, while others need to be modified. The mechanistic model
53 they introduced is called the SIR model, for Susceptible-Infected-
54 Recovered/Removed: A susceptible population is infected by the introduction of a
55 few infected individuals. The infected population eventually recovered, and in the
56 process acquiring immunity to the original infectious disease, or dead (removed).
57 An epidemic ends when susceptibles are exhausted as most of the population is
58 immunized this way, achieving “herd immunity”. There have been many variants to
59 this basic model. One common modification is to add an extra population of E for
60 exposed individuals who are not yet infectious, in the so-called SEIR model. Other---
61 agent-based---- models take advantage of the modern computing power to further
62 subdivide the population into many subgroups and even simulate movements of
63 individuals. But basic concepts are similar to the SIR model. These mechanistic
64 models are of critical importance before the outbreak starts, for the models could be
65 used to explore various scenarios for policy decisions on social distancing and
66 lockdown. Once the outbreak starts, a different kind of models are needed----so far
67 not well developed----which are capable of using real-time data to provide the
68 needed parameters for epidemic management, predict various turning points and
69 peak medical resource needs, and monitor the effect of compliance of the quarantine.
70 In this paper, we introduce such a statistical model, supported by theory and
71 validated by observed data.

72

73 There are actually two possible end states of an epidemic: one is through “herd
74 immunity” mentioned above when we discuss the mechanistic models, and the
75 other is an unstable state achieved by suppressing contacts among individuals,
76 called “suppressed equilibrium” here, which is pursued by most countries in this
77 pandemic, though many policy makers may not be aware of the distinction between
78 the two. This second state is (parametrically) unstable because if the social
79 distancing measures are relaxed and the businesses reopened, the disease could
80 initiate a second wave, as most of the population has not acquired immunity. Even if
81 the epidemic ends in one country, there could still be subsequent waves of infection
82 by imports from abroad unless there is strict quarantine of cross-border travelers.
83 There is also a third route: to suppress just enough so that the maximum cases of
84 hospitalizations are kept below the maximum capacity of the medical system in each
85 district, delaying long enough either for the population to eventually achieve herd
86 immunity or for a vaccine to be developed for the disease. More people would be
87 infected and more would die in this approach than the suppressed equilibrium, but
88 less than that in the herd immunity scenario, where the hospitals may become

89 overwhelmed. Which approach to take is the difficult decision confronting policy
90 makers at the beginning of an epidemic.

91
92 It was reported (2) that UK first contemplated not suppressing the epidemic
93 through lockdowns, fearing that doing so would only lead to a larger second
94 outbreak because most of the population would not have gained immunity. So the
95 plan was to let the epidemic run its course while protecting the elderly. But when
96 shown a model prediction (4) that such a “do-nothing scenario” would lead to
97 500,000 deaths and 81% of the population infected, policy makers changed course
98 and imposed strict counter measures. This is an important and proper role for a
99 model, to prompt policy actions to combat the spread of the disease. Once the
100 outbreak started, the accuracy of the mechanistic model predictions cannot be
101 verified as the forecast forever changed the course of the epidemic in UK. Health
102 officials in Sweden didn’t believe in models and decided to pursue “herd immunity”
103 starting 12 March.

104
105 The number of people that will have to be infected before achieving herd immunity
106 depends on how contagious the disease is. The 16 March report of Ferguson et al.(4)
107 assumed an infection rate, expressed in terms of Basis Reproductive Number R_0 of
108 2.4. For US it predicted that 81% of the population would have to be infected in this
109 do-nothing scenario, or about 250 million, resulting in 2.2 million dead. Later
110 updates in the 30 March report of Flaxman et al. (5) suggested that R_0 should be
111 above 4 for European countries studied. Estimating this number will be one of the
112 tasks in the present work so that one can evaluate what it entails for the herd-
113 immunity approach. We will later show, directly from data, that this estimate of R_0
114 ~ 4 also holds for the US, and in fact approximately so for every country we
115 examined. So over 90% of the US population would need to be infected before herd
116 immunity could be achieved. COVID-19 turns out to be much more contagious than
117 originally thought. See also (6).

118
119 Since a “suppressed equilibrium” is achieved in a very different manner than that for
120 the “herd immunity”, our estimate of the end date of the epidemic as a consequence
121 of contact suppression is not based on the number of susceptibles, S , approaching a
122 small critical value (i.e. when most of the population is infected, hence acquiring
123 immunity), but the daily new cases $N(t)$ approaching zero and remaining so for two
124 incubation periods. For prediction purpose, the date when the $N(t)$ is near zero is
125 estimated by 3 standard deviations from its peak. These two quantities, the peak
126 and the standard deviation, can be extracted from the data as the epidemic is
127 developing. Our estimate of the end of the epidemic is earlier, usually significantly
128 so, because it does not depend on a high percentage of the population having been
129 infected to achieve herd immunity.

130
131 For South Korea, the epidemic in that country ended with just 0.02% of its
132 population infected. Wuhan, the epidemic ended with less than 0.5% of its
133 population infected, both less than 1% of that required to achieve herd immunity as
134 predicted by most mechanistic models. Even if these numbers are multiplied by a

135 factor of 10 (as some in the media suggested that the Wuhan data may be an
136 underestimate), it is still one order of magnitude less. Furthermore, it is generally
137 accepted that South Korea's data are good. Indeed, the situation in these countries
138 represents early examples of the "suppressed equilibrium". Because of its much
139 lower number of deaths, such an end state is a goal that most countries have
140 decided to pursue, despite the enormous toll on the economy due to the much
141 reduced business activity for the two to three months that it would take to achieve it.

142
143 The above-quoted numbers for Wuhan and South Korea are for the confirmed cases,
144 and do not include the asymptomatic infected. Recently (in early April) about 3,330
145 individuals in Santa Clara County in California were tested for antibodies to the
146 COVID-19 virus in their blood (7). 50 were tested positive (meaning that they were
147 exposed in the past to the virus), yielding a crude prevalence rate of 1.5%. When
148 weighed by demographics and extrapolated statistically to the whole county's
149 population, it was calculated that 2.8% of the population could have been infected,
150 with a 95% confidence range of 1.3-4.7%. These numbers, less than 5%, are also
151 much less than what is required to achieve herd immunity, in a county where the
152 epidemic is waning at the time.

153
154 In pursuing the "suppressed equilibrium" and monitoring the progress towards it, a
155 second type of statistical models is required. This type of models should reflect the
156 contact-reduction measures already in place and the degree of compliance by the
157 population in each region. Unlike China and South Korea, many European countries
158 imposed these measures in stages, and so the contact rate among the population
159 was reduced in a time-dependent way. In US, even before the epidemic is over, some
160 states are starting to reopen businesses. Only a real-time, data driven model can
161 reflect these changes. Such a model should be based on sound epidemiological
162 principles. Prediction based on a purely statistical model without an epidemiological
163 foundation, even though using real-time and past data, is akin to watching the daily
164 stock market fluctuations and performing "technical analysis" to ask when the peak
165 is for an investor to time a sell order. A prediction could be made but the
166 uncertainty would be so large that it is just likely for the prediction to be right as for
167 it to be wrong.

168
169 Each type of models has its strengths and weaknesses. For the mechanistic model,
170 such as SIR and SEIR or agent-based versions, a key parameter, the infection rate, is
171 not known for an emerging disease such as SARS-CoV-2, and this has been a source
172 of difficulty with predictions using such models. For the second type of models,
173 especially the purely statistical models without epidemiological basis, it is not
174 known which quantity of the epidemic is predictable. For example, there have been
175 many empirical models based on the assumption that the progression of daily cases
176 follow a Gaussian "epidemic curve" in time, starting with the early model of William
177 Farr in 1840: "Law of Epidemics" in his second annual report to the Registrar
178 General of England and Wales (8). Lacking the epidemiological mechanism that
179 Kermack and McKendrick (3) later proposed, the "law" simply reflected Farr's
180 conviction that the observed deceleration of the rate of increase of infected would

181 not lead to an impending catastrophe but to a crest and then accelerated decline.
182 The latest is that of the Institute of Health Metrics and Evaluation (IHME)(9). It
183 turns out that fitting three parameters that define a Gaussian to a short time series
184 and then using that Gaussian to predict the peaks of the epidemic is an ill-posed
185 problem(10). The uncertainty for next-day prediction is near 100%, and should
186 increase further for predictions a few days out (11). In this work we pay special
187 attention to which epidemiological property is predictable, and we quantify the
188 uncertainty of our prediction.

189
190 The search for the correct “geometry of epidemic curves” has a long history in
191 statistical modelling. Farr’s law is purely descriptive. Farr did not realize that his
192 epidemic curve is Gaussian but nevertheless his descriptive law of second ratios
193 could be used for prediction, though not very accurate. For example (see (12)), if
194 $x_1, x_2, x_3, x_4, x_5, x_6, \dots$ are the successive weekly incidence (i.e. new cases) or mortality,
195 his law says that the ratio of successive ratios of these numbers is a constant:

196
$$\frac{x_4/x_3}{x_2/x_1} = \frac{x_5/x_4}{x_3/x_2} = \dots = K,$$

197 which is less than 1. That is, there is a constant deceleration of the rate of growth of
198 the cases. After measuring this constant from the early weeks’ data, future
199 incidence values can be predicted. It was John Brownlee in 1907 (13) who realized
200 that the above formula, when logarithm is taken----turning the ratio of ratios to
201 difference of differences----is a finite difference form of the second order time-
202 derivative of $\log x$ being a negative constant (12):

203
$$\frac{d^2}{dt^2} \log x = \log K < 0.$$

204 Integrating twice and then taking exponentiation lead to $x(t)$ being a Gaussian form.
205 Brownlee thought this normal form for the epidemic curve is a fundamental law in
206 epidemiology, but his proposed explanation for the declining growth of the
207 incidence of an epidemic as due to decreasing “infectivity” was not well-received by
208 epidemiologists at the time.

209
210 Brownlee (13) provided examples of several epidemics showing that there was fore-
211 aft symmetry in their epidemic curves. For COVID-19, we find that the epidemic
212 curve for Wuhan, China follows a Gaussian, with near fore-aft symmetry, but that for
213 US has a rapid rise but slow decline, definitely not Gaussian. While it may be
214 possible that without human intervention, a solitary outbreak may follow a
215 Gaussian curve, in the modern era of contact suppression, the epidemic curve is
216 shaped by such interventions. We shall explain Wuhan’s shape as due to the fact
217 that the contact suppression measures were consistently imposed throughout the
218 course of the outbreak, while in the case of US, its states and the populace, were
219 relaxing earlier measures on the aft side of the curve, when the new cases declined,
220 creating a fore-aft asymmetry. Therefore, one should take into account that in the
221 modern era, as countries pursue a “suppressed equilibrium” at great economic cost,
222 there is a tendency in countries with decentralized state governments to relax the

223 countermeasures to various degrees once the disease crested, giving rise to
224 subsequent waves of infection. [A second reason is that the US data is an aggregate
225 of data for different epicenters, with staggered recovery. A third reason for the
226 asymmetry may be artificial: In many countries there is usually an increase in the
227 testing of the population as the production of the test kits and facilities to process
228 them ramp up after the initial shortage. Therefore it is to be expected that there
229 would be more cases found later in the course of the epidemic than in the beginning.]

230
231 Brownlee rejected the idea of herd immunity, that the epidemic's decline was due to
232 "an exhaustion of susceptibles, enshrined 20 years later in the SIR model of
233 Kermack and McKendrick (3). His alternative, "infectivity" idea was based on the
234 thinking that the decline was due to "the loss of infecting power on the part of the
235 organism"(13), and that this biological property of the pathogen ("organism")
236 should follow some fundamental law. This biological property of the virus has not
237 been observed in the current COVID-19 pandemic, and does not appear to be a
238 factor. However, Brownlee's idea can be resurrected by modifying the definition of
239 "infectivity" to include social factors, since how many people one infected individual
240 can infect, as measured by the Effective Reproduction Number, R_t , depends on the
241 product of the number of persons contacted during the infectious period and the
242 probability of the contacted person contracting the disease. After implementation of
243 contact-reduction measures, we can actually see from the data (in section 2) the
244 decline of this measure of "infectivity". Furthermore the decline is steeper in
245 countries that have the more stringent contact-reduction policies and
246 implementation. Although both the mechanisms of loss of susceptibles and
247 decrease in "infectivity" are likely at play, with the extremely small percentage of
248 the population infected in the current pandemic, the second mechanism appears to
249 be the dominant one as countries strive to achieve the "suppressed equilibrium".
250 Given this situation, model predictions of the decline of the epidemic based the
251 number of susceptibles decreasing, as in SIR and SEIR models, may be missing the
252 main cause for the observed progression of the disease in the current pandemic.

253
254 Brownlee's idea, with modification expressed above, can be cast in a mathematical
255 form as:

$$256 \quad x_{t+1} = R_t x_t,$$

257 where x_t is the incidence (new cases) at time t , and x_{t+1} is the incidence one
258 infectious period later. An infection period is the duration an infected person
259 remains infectious. R_t is defined earlier as the number of people one infected
260 individual would infect during the period when he is infectious. If R_t is a constant,
261 $R_t = R_0$, the solution to the above finite difference equation is:

$$262 \quad x_t = x_0 (R_0)^t ;$$

263 the solution is an unimpeded exponential growth (since $(R_0)^t = \exp\{t \log R_0\}$) for the
264 relevant case of $R_0 > 1$. Brownlee(14) commented that such an epidemic form is
265 contrary to the facts: "The assumption that the infectivity of an organism is constant,

266 leads to epidemic forms which have no accordance with the actual facts.” With R_t as
267 a decreasing function of time, which we find is actually the case in section 2, the
268 above solution becomes Gaussian-like. Specifically, if R_t decreases by a factor $q < 1$
269 after each period (12, 14), due to a “loss of infecting power”, i.e. $R_t = R_0 q^{(t-1)}$, then the
270 solution is Gaussian:

$$271 \quad x_t = x_0 (R_0)^t q^{t(t-1)/2},$$

272 (since $x_t = x_0 \exp\{t \log R_0 + \frac{1}{2}t(t-1)\log q\}$, noting $\log q < 0$).

273 The exponential growth from $(R_0)^t$ is eventually overtaken by the more rapid
274 decrease of q^{t^2} . Note that in this argument, no mention is made of the decrease of
275 the number of susceptibles; this is not needed when the decrease is so small
276 compared to the population as a whole.

277
278 This paper is organized as follows: We first introduce the relationship between the
279 net infection rate and the Reproduction Numbers. The epidemiological basis for our
280 model is discussed. Then, we will present a suite of prediction tools for epidemic
281 management. We will also provide a summary of the COVID-19 inferred
282 epidemiological characteristics for various countries in Asia, Europe and US. Finally,
283 a discussion and conclusion will be given. The theoretical support is given in the
284 appendix.

285

286 **2. The net infection rate and the Reproduction Numbers**

287 Before we discuss our prediction model, we first discuss diagnosing a key parameter
288 used as input in most mechanistic epidemiological models, the infection rate, or
289 equivalently expressed as the *Basic Reproduction Number* or the time-dependent
290 *Effective Reproduction Number*. Using real-time data, we diagnose it for different
291 countries in the world, which actually reflects the underlying influence factors.

292

293 We define in general the *net infection rate* $\alpha(t)$ as the time-varying exponential
294 growth rate of the active infected cases (15):

$$295 \quad \alpha(t) = \frac{d}{dt} \log I(t) = \frac{\frac{d}{dt} I(t)}{I(t)}.$$

296 The active infected number, $I(t)$ is given by the equation that describes its rate of
297 increase as the daily new infected cases $N(t)$ minus the daily recovered/removed
298 case, $R(t)$:

$$299 \quad \frac{d}{dt} I = N(t) - R(t).$$

300 The dead (“removed”) is included in $R(t)$ in our calculations. The peak number of
301 active infected cases is a key parameter in the planning for hospital resources. This
302 turning point, denoted by t_p , can be located in a local-in-time manner by when R
303 starting to exceed N , without first accumulate the data in time to find $I(t)$. Maximum

304 demand for hospital resources occurs at its peak, and not at the peak of $N(t)$,
305 although the latter is a more commonly reported quantity. $R(t)$ is not a factor in the
306 initial rise of the outbreak, nor is it needed to explain why the rise slows and
307 eventually crests. However in the waning phase, when the new cases are smaller,
308 the rising recovered cases need to be taken into account to determine t_p .

309
310 In traditional mechanistic models, such as the SIR model(3), there is also a time-
311 dependent net infection rate, which at $t=0$, when the population is completely
312 susceptible, is related to the *Basic Reproduction Number* R_0 . See ref (16) for a
313 discussion of the complexities associated with this key parameter. We will not be
314 using the SIR model but it is useful to relate our general definition to what is
315 traditionally used. The equation for I in the SIR model is:

316
317
$$dI/dt = aSI - bI = bI(aS/b - 1),$$

318 where $aS(t)$ is the infection rate and b is the recovery/removal rate.

319 Therefore
$$\alpha(t) = \frac{dI/dt}{I} = b\left(\frac{aS(t)}{b} - 1\right) = b(R_t - 1).$$

320 So for the SIR model $R_t = \frac{aS(t)}{b}$. So as the population of the susceptible $S(t)$

321 decreases---“exhaustion of the susceptibles” as hosts for the disease----the Effective
322 Reproduction Number decreases. More relevant for our discussion is the possibility
323 that the infection rate a could change as a result of contact reduction measures in
324 place.

325
326 Initially when the whole population is not yet infected, the Basic Reproduction

327 Number is
$$R_0 = \frac{aS(0)}{b} = R_t(0) = \alpha(0)b + 1.$$

328
329 The above-defined time-dependent net infection rate generalizes this concept to be
330 independent of the SIR or other models: If in the course of an epidemic, $\alpha(t)$ is
331 positive, the number of infectives will grow exponentially, reaching a peak number
332 of infectives when $\alpha(t)=0$ at $t=t_p$, which is a critical turning point mentioned
333 above. Then the total number of active infectives will decrease exponentially. In
334 terms of $R_t = \alpha(t)/b + 1$, if this number is greater (less) than 1 the total number of
335 active infectives will grow (decrease) at time t . We will here use $\alpha(t)$ directly. R_t
336 however is the more watched quantity by the mainstream modelers (16). It can be
337 calculated from the net infection rate, but will require a parameter b , the recovery
338 rate, which may be different for different regions. Furthermore, many countries do
339 not keep adequate records on those who recovered, and so there is an uncertainty in
340 estimating b . In Figure 2, R_t is obtained by estimating this parameter as

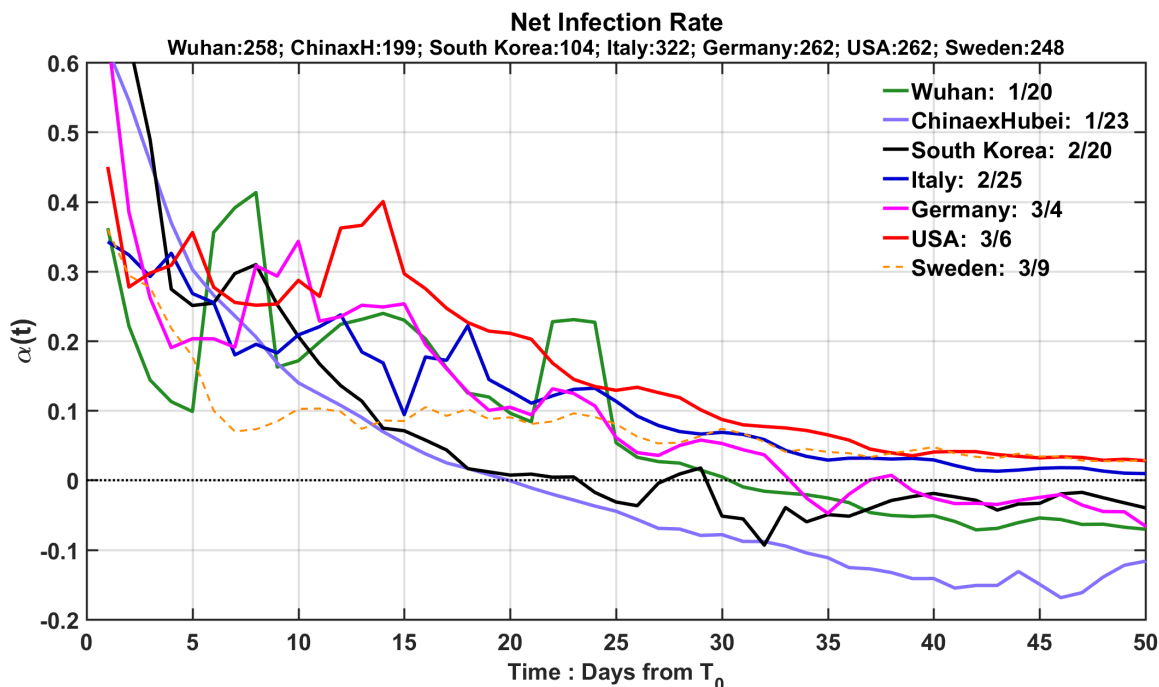
341 $b \approx 1/\sigma_R \approx 1/\sigma_N$, where σ_R is the standard deviation for the distribution of the
342 daily recovered and σ_N is that for the daily newly infected numbers. These

343 parameters are calculated for each country (see Table 1). R_0 is obtained from R_t in
344 the initial period, before there is significant recovered population.

345
346 For an emerging disease such as SARS-CoV-2, there is not enough statistics for
347 estimating the true infection rate (e.g. the parameter a in the SIR model). Models
348 usually assume a value of R_0 or perform scenario calculations for a range of values of
349 R_0 .

350
351 Figure 1 shows the net infection rate for several countries. Since the official data
352 that we use include only the confirmed cases (“cases” for short), and these tend to
353 have more serious symptoms that require hospitalization, the reported I cases are
354 commonly referred to as *total hospitalizations*. The peak of total hospitalizations is
355 closely watched by hospital administrators and policy makers. $\alpha(t)$ is commonly
356 referred to as *rate of hospitalization*. Its inverse gives the e-folding time in days for
357 the cases in an outbreak. A value, of say α between 0.3 and 0.4, where most
358 countries cluster in the initial period, implies an e-folding time of about 3 days
359 (doubling time of 2 days). The much higher values α for many regions in the
360 beginning of our data record may not be due to indigenous disease infection. See
361 later discussion.

362
363



364
365 **Figure 1.** The time-dependent net infection rate (in units of 1/day) as a function of
366 time starting on the date (listed in the inset) when the newly confirmed case
367 number first exceeds 100 for each region. To obtain the actual calendar date, add
368 the dates on the horizontal axis to the starting date indicated in the inserted legend.
369 The number of confirmed cases on the starting date is listed at the top. Three day
370 averaging on the raw data has been used.

371

372 The time for different countries is aligned in Figure 1 to begin the time series when
373 each region reached 100 new cases. This way, the progression of the epidemic in
374 each country can be compared. Figure 1 reveals the effects of different policy
375 measures each country adopted. First, South Korea and China exHubei have similar
376 net infection rates (until past their respective turning point); both are much lower
377 than other countries. In the case of South Korea, the government identified early
378 that its epicenter of the epidemic was at church gatherings in the city of Daegu and
379 North Gyeongsang province, where 90% of the initial cases were found. Specifically,
380 a confirmed COVID-19 patient was reported to have attended the Shincheonji
381 Church of Jesus services twice on February 9th and 16th. Then aggressive contact
382 tracing was pursued. After the turning point, South Korea soon experienced some
383 second wave episodes, which were successfully contained. These two regions'
384 rigorously implemented contact reduction and aggressive pursuit of 'test-trace-treat'
385 measures led to them being the extreme examples of the "suppressed equilibrium".

386

387 Germany and Italy have similar exponential growth rates of the net infected case
388 numbers, both slightly higher than Wuhan. More surprisingly, US has the highest
389 exponential net infection rate, 1.5 times that of Germany and Italy and twice that of
390 Wuhan. This can be attributed to the fact that US so far does not have a nation-wide
391 lockdown, and Europe has had partial lockdowns in phases. Germany took a week
392 longer than Wuhan to reach its turning point, while US will take weeks longer than
393 Germany. China outside Hubei reached its turning point early, in fact 9 days earlier
394 than the epicenter, Wuhan. This fact is significant, for it is qualitatively different
395 than many mechanistic model predictions, which had the epicenter achieving its
396 turning point 1-2 weeks earlier than China outside Hubei (17), probably based on
397 the herd immunity concept.

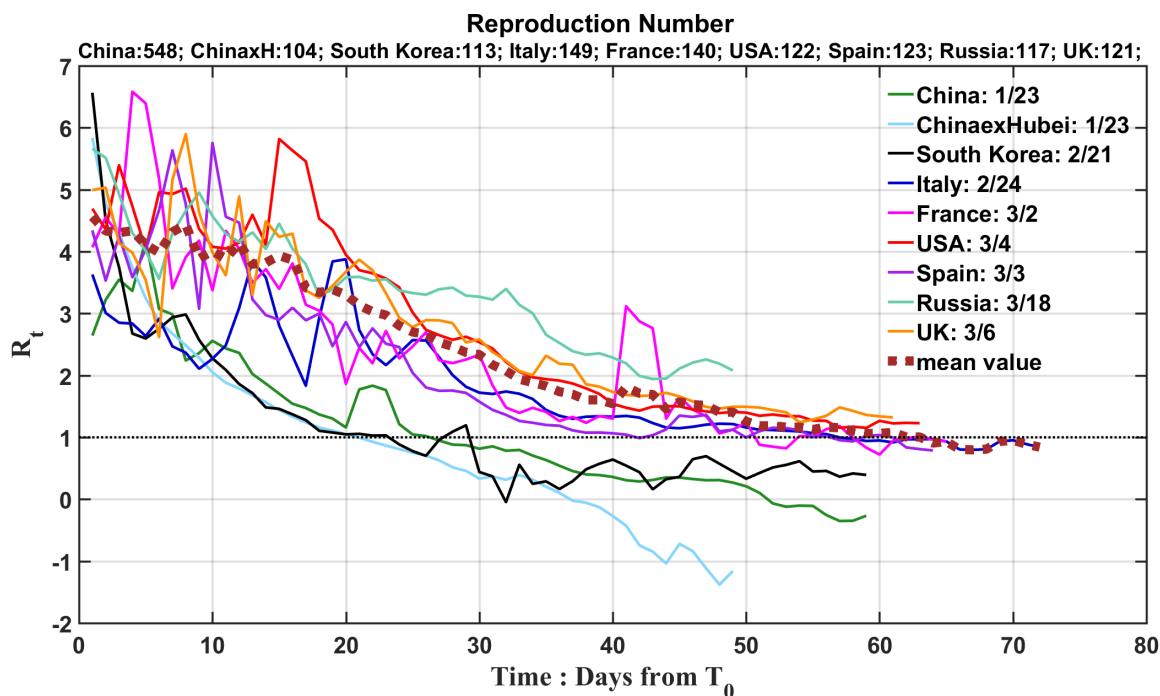
398

399 The net infection rates for China outside Hubei and South Korea are more
400 monotonic than other regions shown. This is due to the fact that there was not a
401 piece-meal imposition of social-distancing measures, unlike other western countries.
402 The strict measures were imposed and enforced throughout the course. [For
403 Wuhan, the large spike on day 23(12 February) was due to a change in the
404 diagnostic criteria from a positive nucleic acid test to chest scans, booking in one
405 day more than 13,000 cases.]

406

407 The case of Sweden needs a special explanation. The epidemic in Sweden initially
408 grows with an e-folding time of around 3 days, in line with other countries. Then on
409 12 March the government announced that because of limited resources, it no longer
410 would test for the COVID-19 infection, except for those with serious symptoms
411 already in the hospitals who furthermore are also in the high-risk group. As a result
412 the new cases took a nose-dive on that day, leading to an artificially low net
413 infection rate of 0.1, implying a 10 day e-folding time. The denominator in the
414 calculation for $\alpha(t)$ is $I(t)$, which is an accumulated quantity, and includes those
415 who tested positive prior to 12 March under more liberal criteria, and so this

416 situation leads to a flat, low level of $\alpha(t)$ just above 0. It would eventually cross 0
 417 with a large enough death numbers. Sweden’s policy decision to pursue “herd
 418 immunity” (while protecting the elderly) has been touted as a viable and perhaps
 419 preferable approach to those of other countries in their pursuit of “suppressed
 420 equilibrium”. It only encouraged those over 70 to stay home and banned visits to the
 421 nursing home and gatherings with over 50 people, while business, stores,
 422 restaurants and kindergarten through grade nine are open. The success or failure of
 423 this approach cannot be evaluated by the incomplete data. Intriguingly, based on
 424 the recorded death number, it shows that Sweden’s toll is 5 and 11 times that of its
 425 neighbors Denmark and Norway, respectively. The population in each of these
 426 neighbors is half that of Sweden.
 427



428
 429
 430

431 **Figure 2.** Effective Reproduction Number for each country or region. The horizontal
 432 axis denotes days since day 0 (the corresponding calendar date is given in the inset),
 433 which is the starting date for our calculation. This date is determined by the
 434 threshold that the accumulated number of infectives first exceeds 100. The actual
 435 number for each region on that day is listed at the top. The thick dashed curve is
 436 the average of the curves for USA and European countries, including Russia.
 437

438 Figure 2 converts Figure 1 to show R_t for each country. It shows that R_t clusters
 439 around 4 for all countries in the initial period. Because of the problems for the data
 440 in the initial period, the curves cannot be extended back in time to deduce R_0 . But
 441 based on R_t a few days later, R_0 for COVID-19 should be around 4, similar to that for
 442 SARS. It was originally thought that COVID-19’s R_0 was between 2.0 to 2.5 (18),
 443 seemingly much less contagious than SARS at 4 (19). It is also much more

444 contagious than the 2009 swine flu pandemic, caused by the H1N1 virus, whose R_0
445 was estimated(20) to be 1.4 to 1.6.

446

447 In deducing the Reproduction Numbers we should not count the large spike for
448 China outside Hubei on day 4. That increase was not due to indigenous
449 transmission, for most of the initial cases were imported from Hubei. This should
450 not be used to infer the Reproduction Number. Similarly for South Korea in the first
451 few days shown.

452

453 As is in Figure 1, the decrease of R_t from 4 to 1 for different countries reflects
454 different level of contact-reduction measures adopted and enforced. With China
455 outside Hubei and South Korea sloping more steeply than Europe than US. The
456 behavior of these numbers for the European countries are rather similar to the
457 model results of Flaxman et al.(5) of Imperial College (their Figure 2), who imposed
458 these measures in their model on the dates they were actually imposed. With their
459 emphasis on more accurately predicting the mortality, when $R_0 \sim 4$ is used in their
460 SEIR model the modeled death numbers are close to the reported deaths, but the
461 number of infected in the model is an order of magnitude higher (Figure 2 in (5)).
462 This is reasonable and consistent with our earlier statement that the data of infected
463 cases do not include the asymptomatics or those with mild symptoms who were not
464 tested, but reflect those with more serious symptoms that required hospitalization,
465 and therefore more prone to die.

466

467 UK's record for the recovered is almost nonexistent, and what is available shows
468 that the recovered number is only a few percent of the deaths, which does not
469 appear to be reasonable. Without the recovered in the data, UK's R_t hovers above 1.
470 There appears to be a similar situation in some other countries, such as Italy.
471 Therefore, the behavior of R_t in the later stages of the epidemic (in the
472 neighborhood of the turning point) is probably not correctly depicted by the data
473 shown for these countries. Nevertheless, in the initial period, when the number of
474 recovered is small, the data shown can be used to estimate R_t and R_0 .

475

476 SEIR model was also used to deduce the Reproduction Number by Institute of
477 Disease Modeling (21) in an effort to monitor the effect of social distancing
478 measures adopted near Seattle (King County, Washington). They found that R_t was
479 reduced from 2.7 to 1.4. Since it was not below 1, the Institute's report advised
480 continuing the measures in place. Since the report rate p , that is, the ratio of the
481 number of reported cases vs. the true infected number, was unknown, the authors
482 assumed a range of values and obtained, $R_0 \sim 2.7 \pm 0.9$. One can see from this
483 application of the Reproduction Numbers how important it is to monitor in real-
484 time the progress of policy measures to determine whether it is time to relax the
485 measures in place. And also how difficult it is to infer these numbers.

486

487 The report rate p is the ratio of the confirmed cases to the true infected numbers. If
488 it were known, the case numbers can be divided by p to yield the true infected

489 numbers. Because it is largely unknown at the present time, it creates the
490 aforementioned uncertainty in estimating Reproduction Numbers using models.
491 However, since our net infection rate $\alpha(t)$ is a ratio of the derivative of I and I ,
492 dividing each by p does not change the ratio. We therefore recommend using the
493 net infection rate for real-time monitoring instead of the Reproduction Number.
494 The assumption is that the report rate has not changed in time, at least not during
495 the two-week period prior to t that the accumulation takes place for the
496 denominator (see Appendix). This same consideration is also partly behind our later
497 using the ratio of new cases to recovered cases for prediction.

498
499 The problem with the data for Sweden is that in addition to the change in the report
500 rate, the record for the recovered is not available.

501
502 This discussion also highlights the difficulty of using mechanistic models for real-
503 time monitoring. In addition to not knowing the report rate p to compare model
504 output with the reported case numbers, a key parameter needed in the mechanistic
505 models, the infection rate a , is largely unknown for the emerging disease. There is a
506 large population of asymptomatic, untested and unreported infectives. This infected
507 population is nevertheless infectious and produces some of the infected cases
508 reported. It would usually lead to an overestimate of the infection rate a when only
509 the reported cases are used in the estimation. Our tools do not need to use an
510 infection rate for prediction.

511 512 **3. Epidemiological basis**

513 After establishing the relationship between the net infection rate and the
514 Reproduction Numbers, we briefly summarize the main ingredients to establish the
515 epidemiological basis of our model. Details of our model are given in the Appendix.
516 The model is used here only to infer general properties of an outbreak, and to
517 discover which properties can be predicted.

518
519 Based on epidemiological theory, the infected population is governed by the Von
520 Foerster equation in an age-structure population model (see (22)), where “age”, s , is
521 days since first infected. The infection at age 0 is governed by the “birth” dynamics,
522 the infection dynamics, and death due to natural causes is ignored. An infected
523 individual is assumed to be infectious in an age-dependent way until T , when the
524 individual is either cured or dead, in either case no longer belonging to the
525 population of the infectives, where T is the mean recovery/removal period. The
526 solution to this partial equation (via the method of characteristics) yields certain
527 rather general results.

528
529 **Conservation Law.** For $t > T$

530
531
$$R(t) = N(t - T).$$

532 The distribution of the newly recovery/removal follows that of the newly infected
533 with a time delay of T . When applied to hospitalizations, it says that those who are

534 admitted to the hospital either recover after a mean hospital stay of T days, or dead
535 after a similar number of days. A more complicated relationship holds for $t < T$.

536

537 Since hospital stay acts like a filter for $R(t)$, the profile for R is slightly wider than
538 that for N . Appendix takes this into account.

539

540 $R(t)$ here actually consists of two parts: the recovered individuals $R_c(t)$ and the
541 removed (dead) individuals, $D(t)$: $R(t) = R_c(t) + D(t)$.

542

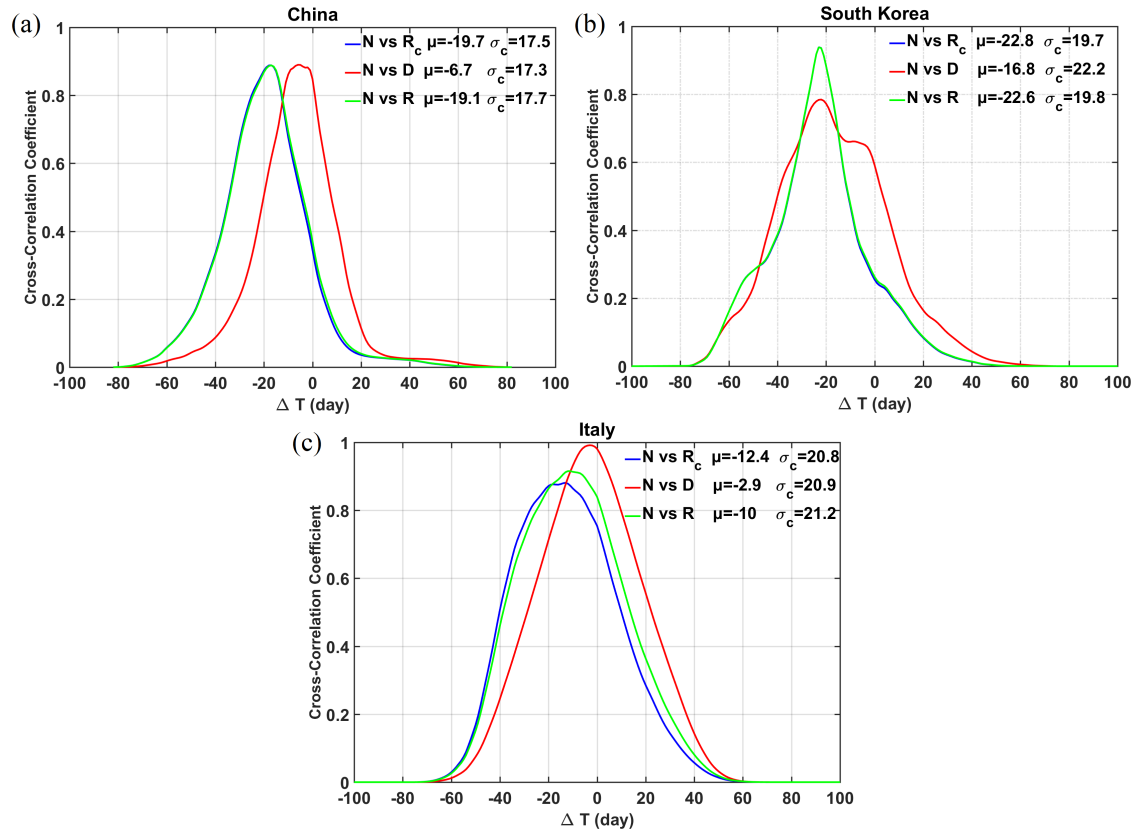
543 **Validation after the fact:** This fundamental relationship can be validated
544 statistically with data, provided the data is long enough. This is one of the ways the
545 mean recovery period T is determined statistically from data, but it is not practical
546 in the early phase of the epidemic.

547

548 Since $R_c(t)$ and $D(t)$ are a subset of $R(t)$, the time lag relationship should also hold.
549 Additional insight can be gained by looking at them separately. Figures 3, obtained
550 using the longest data from China, South Korea and Italy during the COVID-19
551 pandemic, shows that $N(t)$ and $R(t)$ are highly correlated: with correlation
552 coefficients all over 0.9 when both distributions are smoothed with 3-point boxcar.
553 The mean time delay of the correlation is denoted by μ , which can be interpreted as
554 a statistical mean of T . The spread of the cross-correlation is measured by its
555 standard deviation, σ_c ; its value is also given in the figure. The lag time (and
556 standard deviation) of $R(t)$ for China is 19 days ($\sigma_c=17.7$ days) for South Korea is 23
557 days ($\sigma_c=19.8$ days) and for Italy is only 10 days ($\sigma_c=21.2$ days). Due to their low Case
558 Fatality Rate (CFR), there is practically no difference between $R(t)$ and $R_c(t)$ for
559 China and Korea. But for Italy, which has a high CFR, there are differences among
560 the constituting parts. Why Italy's lag time between $R(t)$ and $N(t)$ is shorter than
561 China and South Korea needs a deeper examination, and it does not necessarily
562 mean the shorter the better. The lag time of $D(t)$ for China is 7 days, for Korea is 17
563 and for Italy is only 3 days. The short survival time for patients in the hospital is an
564 indication that the Italian hospitals were overwhelmed and allocation of ventilators
565 was selective to those more likely to survive. This mortality component reduces the
566 overall time for recovery/removal for Italy to 10 days.

567

568



569
570

571 **Figure 3.** Lagged cross-correlation of $R_c(t)$, $D(t)$ and $R(t)$ with $N(t)$ for China, South
572 Korea and Italy. Collectively or individually, the conservation law and time lag
573 relationship are validated. The separated $R_c(t)$ and $D(t)$ reveal the stress the
574 regional hospitals had experienced.

575

576 Epidemic curve.

577 Furthermore, the solution to the model equations (in Appendix) shows that the
578 time-dependent profile of $N(t)$, sometimes referred to as “the epidemic curve”, is
579 determined by the “age” distribution, and the method of characteristics in the
580 solution converts that “age” distribution into a time profile. The details of the “age”
581 distribution depend on the “birth” process and on how long it has been since the
582 outbreak first begins. We assume that we do not mix data from different epicenters
583 (the homogeneity assumption). Focusing on a single outbreak, which starts at $t=0$,
584 we examine the “age” distribution at a time t_B much longer than the initial
585 incubation period. Then we assume there exists a full spectrum of age with s
586 between 0 and T in a Gaussian-like form. To the right of the peak in the “age”
587 distribution, it is easy to understand that those who are “older” should be less in
588 number because they were infected during an earlier stage of the epidemic. As the
589 epidemic grows exponentially, there are more and more “younger” infectives. To the
590 left of the peak in “age” distribution many infectives are at various stages of
591 incubation. Those who are newly infected may be much large in number, but since
592 they have not yet developed symptoms, they are less likely to be tested, hospitalized,

593 and contribute to the “case” record. For Covid-19, the peak infectiousness occurs in
594 a period just before and after the onset of symptoms, around 5-6 days since first
595 infected(23). The above discussion gives an epidemiological justification for a
596 Gaussian-like distribution of the “age” distribution of the “new cases”, more
597 appropriately called “new hospitalization per day”. It is consistent with the
598 argument given in the Introduction in favor of a Gaussian-like epidemiological curve.
599 It should be pointed out that historically, for obvious reasons, asymptomatics were
600 also not included in the “incidence”, or “mortality”, which were used by Farr to
601 obtain his law of second ratios or by Brownlee to obtain his normal curves.

602
603 With the possible exception of South Korea, in most countries there could be
604 multiple seeding of the outbreak occurring at slightly different times. These
605 staggered series of outbreaks merge in a continuum in the data for that country,
606 leading to a standard deviation in the new cases that is wider than that from the
607 “age” distribution. This is a complication we need to take into account.

608
609 Also, in many countries, pressure mounts for policy makers to relax the contact-
610 reduction measures when case counts pass the peak and are declining. In some
611 countries where the restrictions are gradually lifted, we should expect a long tail in
612 the epidemic profile, which is therefore not symmetric with respect to the peak.
613 This external influence to the original expected progression of the course should be
614 monitored and adjustment to predictions made in real time. As will be discussed in
615 subsequent sections, our model has the ability to adapt to these changes in its role
616 as a monitoring tool. Nevertheless, we find, after the fact, that predictions for the
617 various turning points are still accurate even without taking into account the
618 changes. It turns out that, consistent with the above discussion, the relaxation of
619 contact-reduction measures, which lengthens the standard deviation of the new
620 cases, is only significant in the later stages of the course, and can be ignored before
621 the peak, but the prediction on the evolution after the peaks on quantities such as
622 the end of the epidemic and the total number of infected is likely not accurate unless
623 the changes are taken into account. (This has not been done in Table 1.)

624 625 **4. A suite of tools for tracking the epidemic.**

626 Based on theoretical considerations, we developed a suite of tools for monitoring
627 the evolution of the epidemic. They can be used to predict the timing of several
628 turning points and the number of infectives associated with each. These include t_N ,
629 the peak of the daily new cases; t_R , the peak of daily recovered/removed cases, and
630 t_p , the turning point where the active infected cases (AIC), or total hospitalizations,
631 is a maximum. This is the point of maximum strain on the hospital resources. It
632 needs to be closely monitored to keep the AIC below the hospital capacity. We also
633 estimate the date when the epidemic ends and the country can be reopened, t_c .

634
635 A more accurate and robust prediction tool is based on the ratio of $N(t)$ and $R(t)$,
636 designated as the NR ratio. This ratio also alleviates to some extent the problem

637 related to the data of reported cases being a fraction p of the true numbers, as p
638 cancels out in the ratio. Unfortunately, some countries, such as UK and Sweden, do
639 not keep adequate record of $R(t)$, and many country do not keep a rigorous standard,
640 which could be detect by the low case recovery rate, indicating the violation, or
641 leakage, of the conservation law. For these countries a less accurate method, in the
642 sense of having larger error bars, can still be used when only information on $N(t)$ is
643 available. These are described below. We have used the data from countries that
644 have the longest records for COVID-19 to verify properties of these methods.

645

646 **4.1 Log of NR ratio:**

647

648 We define the NR ratio as

649

$$NR(t)=N(t)/R(t).$$

650

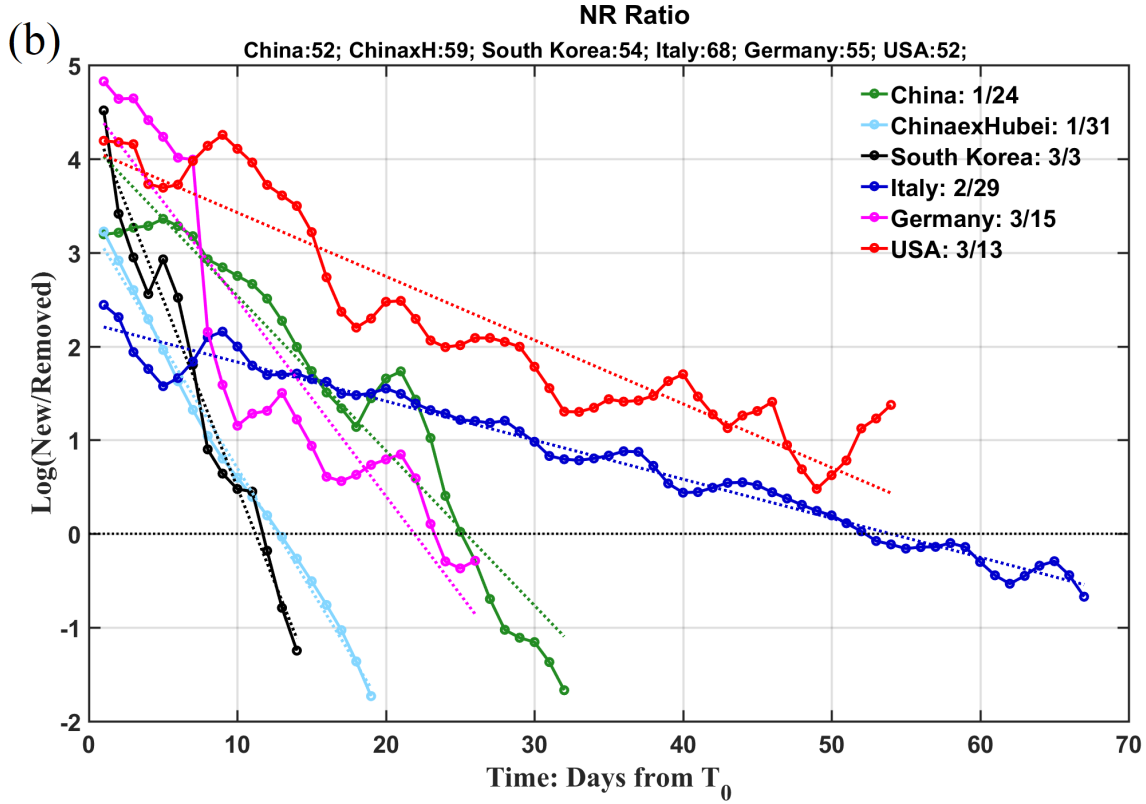
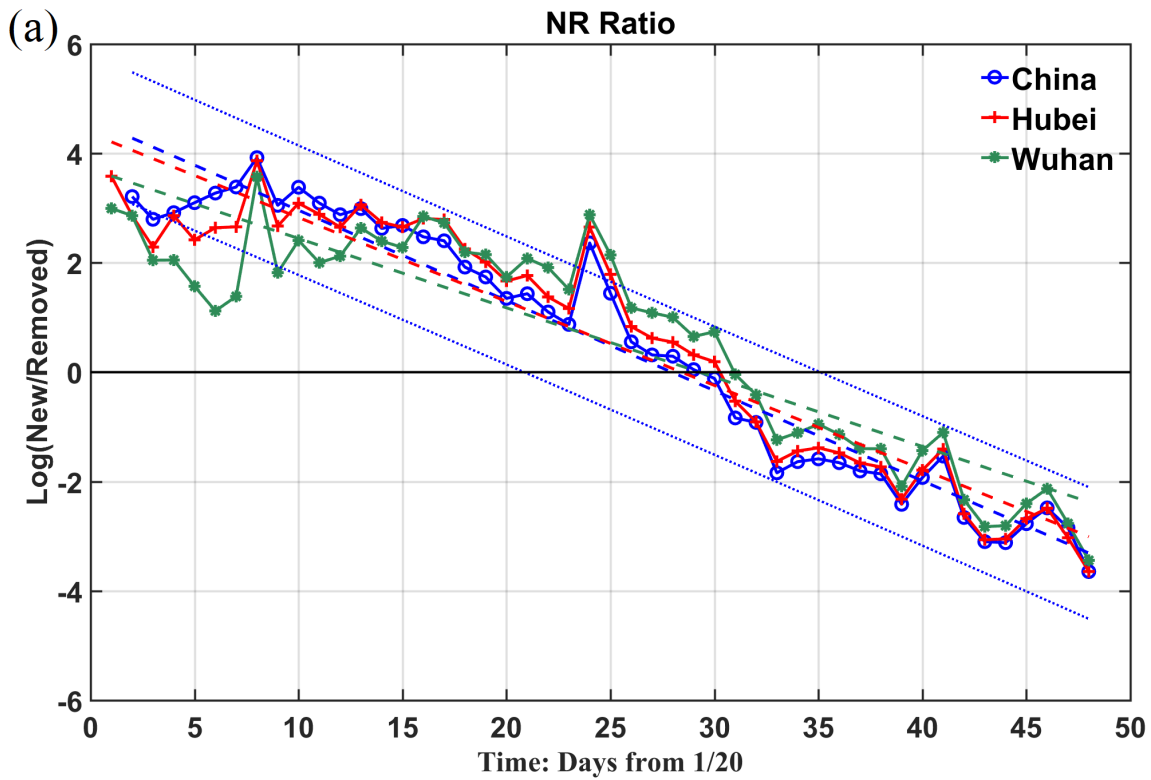
At t_p , $NR=1$.

651

652 We show in Figure 4, using the data of the epidemic for COVID-19 for the longest
653 records available in some countries, that the logarithm of $NR(t)$ lies on a straight line,
654 with small scatter, passing through the turning point t_p . And data for various stages
655 of the epidemic, from the initial exponential growth stage, to near the peak of AIC ,
656 and then past the peak, all lie on the same straight line. The intercept with $\log NR=0$
657 yields the turning point. This line, obtained by linear-least-square fit, is little
658 affected by the rather large artificial spike in the data on 12 February for China,
659 because of its short duration and the logarithmic value. That reporting problem is
660 necessarily of short duration because, on the date of definition change, previous
661 week's cases of infectives according to the new criteria were reported in one day.
662 After that, the book is cleared, and $N(t)$ returned to its normal range.

663

664



665
666

667 **Figure 4.** Natural Logarithm of the ratio of daily newly infected to newly
668 recovered/removed. They lie on straight lines with some small scatter. The straight
669 line obtained by linear-least squares fit is in dotted line with the same color for each
670 country. (a) For China, the slopes of the lines are almost the same but with different
671 intercept; the trend lines cross zero (the black solid line) at different time for
672 different regions indicating different peaking time for *AIC*. The epicenter Wuhan
673 (green) has latest turning point than its province Hubei (pink), which has a later
674 turning point than China as a whole (cyan). As a measure of confidence of the linear-
675 least square fit, the 95% confidence limit for China is given in the figure. Similar
676 confidence limits for other regions have been calculated but are not shown for the
677 sake of clarity of the presentation. (b) Comparing different countries. The time is
678 aligned by plotting the *NR* ratio only when the recovery/removal case numbers first
679 exceed 50 (the actual numbers on that day are listed at the top), and T_0 is that
680 calendar date, listed in the inset. The figure in the bottom panel included the China
681 cases again, to facilitate comparison with other countries, except the data used were
682 smoothed by a 3-point boxcar filter in the lower panel.

683
684 It would be interesting to understand why the empirically determined $\log NR(t)$ lies
685 on a straight line, and what determines its slope. See Appendix for a theoretical
686 support. It is shown that, if the distribution of the new cases is Gaussian-like, the
687 natural logarithm of the *NR* ratio, should be a linear function of time throughout the

688 course of the epidemic. The slope of the line is $-\frac{T}{\sigma_R^2}$.

689 σ_R , is the standard deviation of the recovered case distribution and is close to that of
690 the newly infected case, σ_N , twice of which measures the duration of the epidemic
691 in that region. This expression is still valid even for a sigma that is changing in time
692 as a response to changing social contact measures.

693
694 A comparison of the Logarithm of *NR* ratio for several countries is given in Figure 4
695 (b). A steeper slope is associated with an early turning point, and also a predictor for
696 a shorter duration of the epidemic. The shallowest slopes in Figure 4 (b) was for
697 Italy, where the enormous pressure strained their medical system to the limit,
698 resulting in the largest σ_R value, and one of the highest case fatality rate in the world,
699 at more than 12%. Germany and China have similar slopes. For China outside Hubei,
700 the slope is steepest and the turning point reached 9 days earlier than Wuhan. South
701 Korea's slope is even steeper due to that countries early action. As a result, Italy
702 took a full month longer to reach its turning point than Germany and China, and
703 more than 40 days longer than South Korea.

704 705 **4.2 Predictability**

706 Since the logarithm of *NR* ratio lies on a straight line passing through the turning
707 point of *AIC*, it would be interesting to explore if the turning point, t_p , can be
708 predicted by extrapolation using data weeks before it happened. Extrapolating a
709 straight line is much more practical than other more involved curve fitting

710 algorithms some other groups have adopted. How far in advance this can be done
711 appears to be limited by the poor quality of the initial data, when R is small and
712 highly fluctuating. Figure 5 (a) shows the results of such predictions for China, and 5
713 (b) for Italy. It is a hind cast since the truth is now known (see Figure 8 for how it is
714 determined). The horizontal axis indicates the last date of the data used in the
715 prediction. The beginning date of the data used is 24 January for all experiments for
716 China. Prior to that day, data quality was poor and the newly recovered number
717 was zero in some days, giving an infinite NR ratio. For China outside Hubei, the
718 prediction made on 6 February gives the turning point as 14 February, two days
719 later than the truth. A prediction made on 8 February already converged to the
720 truth of 12 February, and stays near the truth, differing by no more than fractions of
721 a day with more data.

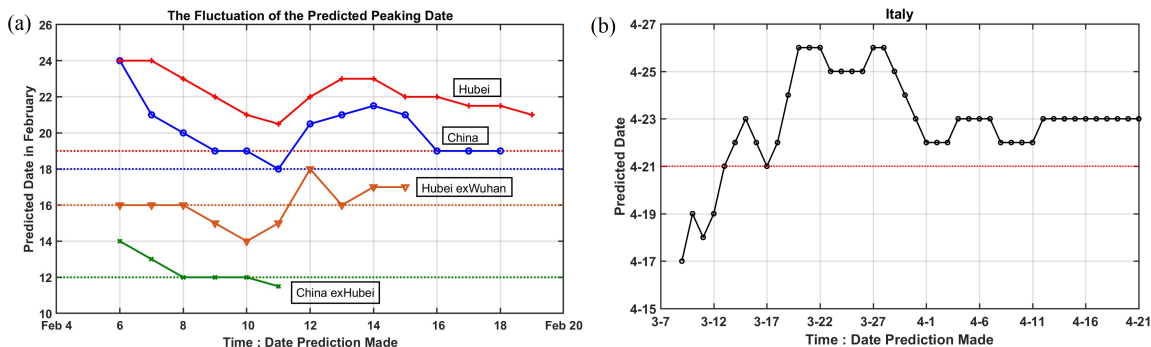
722
723 The huge data glitch on 12 February in Hubei affected the prediction for Hubei, for
724 China as whole, and for Hubei-exWuhan. These three curves all show a bump up
725 starting on 12 February, as the slope of $N(t)$ is artificially lifted. Ironically,
726 predictions made earlier than 12 February are actually better. For example, for
727 China as a whole, predictions made on 9 February and 10 February both give 19
728 February as the turning point, only one day off the truth of 18 February. A
729 prediction made on 11 February actually gives the correct turning point that would
730 occur one week later. At the time these predictions are made, the newly infected
731 cases were rising rapidly, by over 2,000 each day, and later by over 14,000. It would
732 have been incredulous if one were to announce at that time that the epidemic would
733 turn the corner a week later. Even with the huge spike for the regions affected by
734 the Hubei's changing of diagnosis criteria, because of its short duration the artifact
735 affects the predicted value by no more than 3 days, and the prediction accuracy soon
736 recovers for China as a whole. For Hubei, the prediction never converges to the true
737 value, but the over-prediction is only 2 days.

738
739 Uncertainty associated with prediction using this method for China is shown in
740 Figure 4(a). The uncertainty is a few days (the 95% confidence level is ± 5 days)
741 and is usually somewhat larger than the accuracy. A prediction can be found later to
742 be accurate but at the time it was made it may be equally likely for the prediction to
743 be a few days earlier or later. The large uncertainty is again seen to be caused by the
744 1,4000 bump of new cases in one day on 12 February due to the change in definition.
745 For applications to other countries and to future epidemics without a change in the
746 definition of the "infection" to such a large extent, we expect even better prediction
747 accuracy and smaller uncertainty.

748
749 This can be seen in the prediction for Italy. The error of predicting the turning point
750 3 weeks in advance is only 1 or 2 days. In fact a prediction can be made 6 weeks in
751 advance with an accuracy of 5 days or less.

752
753 The prediction for US as a whole is less accurate (with errors up to 10 days) because
754 its data is an aggregate of different epicenters. More accurate predictions can be
755 made by treating each state separately. This is not done here because although the

756 data for new cases and deaths are available for each state, recovered data are not
 757 individually available. It is also not accurate for UK because its data for recovered
 758 may be suspect. See Table 1.
 759
 760
 761



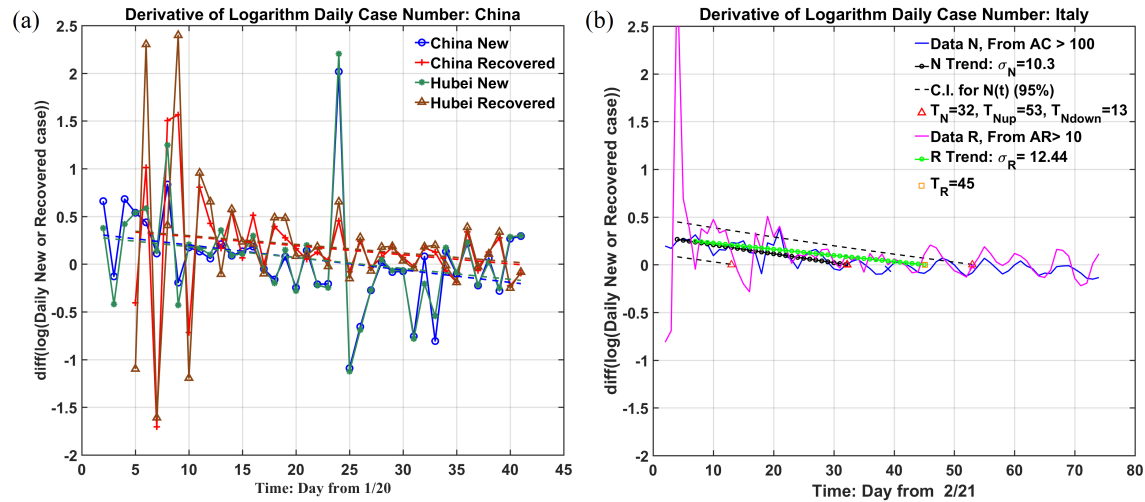
762
 763 **Figure 5.** (a) Prediction of the turning point in *AIC* for different regions of China by
 764 extrapolating the trends in logarithm of *NR*. The horizontal axis indicates the date
 765 the prediction is made using data from 24 January to that date. The vertical axis
 766 gives the dates of the predicted turning point. Dashed horizontal lines indicated the
 767 true dates for the turning point. (b) Prediction of the turning point in *AIC* for Italy.
 768 Data used for all predictions starts on 29 February. The first point shown predicts
 769 the turning point to occur on 19 April (4 days early) 6 weeks in advance using 9
 770 days of data from 29 February to 8 March.

771
 772 **4.3. Derivative of Log of $N(t)$ and $R(t)$.**

773 Interestingly, the derivative of $\log N(t)$ or $\log R(t)$ also lies on a straight line, as
 774 shown in Figure 6 (although the scatter is larger as to be expected for any
 775 differentiation of empirical data). Moreover, the straight line extends without
 776 appreciable change in slope beyond the peak of $N(t)$, suggesting that the distribution
 777 of the newly infected number is approximately Gaussian-like at least up to that point.
 778 For an exponential function, the derivative of its logarithm being a linear function of
 779 time is highly suggestive of a general type of distribution including Gaussian-like or
 780 Rayleigh-like. The recovery time T can be determined as $t_R - t_N$, where t_R is the peak
 781 of $R(t)$ and t_N is the peak of $N(t)$. These two peak times can be obtained by extending
 782 the straight line in Figure 6 (a) to intersect the zero line. This predicted result can be
 783 verified statistically after the fact by the lagged correlation of $R(t)$ and $N(t)$. If the
 784 distribution is indeed Gaussian or even approximately so, the slope in Figure 6
 785 should be proportional to the reciprocal of the square of its standard deviation, σ ,
 786 as:

787
$$\frac{d \log N(t)}{dt} = \frac{-(t - t_N)}{\sigma_N^2}.$$

788
 789 Similarly result holds for the daily number of recovered, $R(t)$.
 790



791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

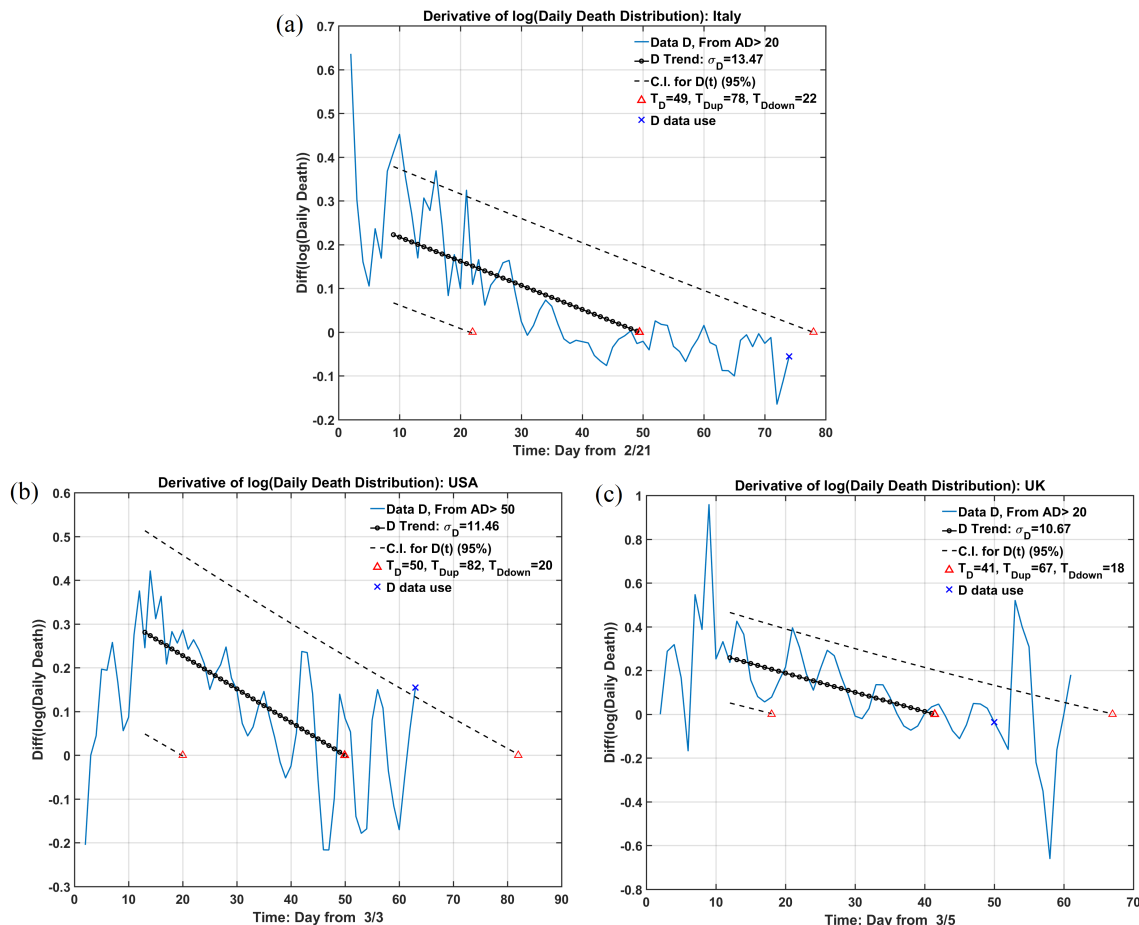
Figure 6 (a). The derivative of the logarithm of daily newly infected or recovered for China and Hubei. Notice the clear separation of the new and recovered cases and also the subtle difference of their slopes. The zero crossings of the trend line give the peak dates of the new and recovered case respectively. And the slopes give an estimate of σ values. (b). Same as (a) but for Italy. Best fit lines for $N(t)$ and $R(t)$ are shown, and 95% confidence limits are shown for $N(t)$ only. Italy's data appear to have a 7-day periodic oscillation, probably caused by a reporting issue. AR indicates accumulated R.

For countries without an adequate record of $R(t)$, this method can still be used for $N(t)$. We can obtain t_N and σ_N . T and t_p cannot be obtained, but can be estimated roughly as $t_p = t_N + T/2$ (see Theory), using $T \sim 20$ days applicable to countries with similar medical systems.

4.4 The number of deaths

Because of public attention, predicting the number of deaths has been the main emphasis for some models. There is also the belief that death numbers are more reliable than case numbers, although that is not necessarily true. Although death number by itself could not satisfy conservation law, death is a subset of the newly confirmed cases; therefore, the death case distribution should follow the new cases with a time delay. In fact, this delay time could provide a measure of the efficacy of the medical system, as explained earlier.

In Figure 7, we present the cases for Italy, USA and UK.



816
 817
 818
 819
 820
 821
 822
 823
 824
 825
 826
 827
 828
 829
 830
 831
 832
 833
 834
 835
 836

Figure 7 (a), (b) and (c). Logarithm for death distribution for Italy, USA and UK. Similar to 4.2, the derivative of logarithm of $D(t)$ follows a straight line approximately, indicating that its distribution is also Gaussian-like, similar to $N(t)$. As a consequence, the peak of $D(t)$ occurs at the intersection of that straight line with zero and could be predicted in advance, although in the three countries shown this peak has already occurred.

5. Inferred epidemiological characteristics.

Table 1 summarizes predictions for USA, some European countries, South Korea and China. In some countries, one or both of the critical points have just occurred. Even so, it is still difficult to know if one is at the peak without using our prediction procedure. The TIC_∞ is the Total Infected Cases for the whole epidemic, calculated as twice the value of TIC at t_N , and AIC_{peak} is the maximum of AIC at t_p . Appendix discusses how these quantities are computed/predicted.

	t_p (r)	t_N (r)	t_D (r)	AIC_{peak-K} (r)	$TIC_{\infty-K}$ (r)	$AFC_{\infty-K}$ (r)	σ_N Days	CFR --%	IPM --Per M	FPM --Per M
China	2-21 (2-17)	2-7 (2-13)	2-17 (2-14)	44.4-54.1 (58.0)	87.7-163 (84.0)	3.8-6.6 (4.6)	8.79	5.53	59.98	3.31
China ex- Hubei	2-11 (2-12)	2-5 (2-4)	2-16 (2-13)	8.99-9.04 (8.94)	15.5-26.1 (15.8)	0.14-0.23 (0.13)	8.49	0.79	11.82	0.09
South Korea	3-15 (3-14)	3-2 (3-1)	3-19 (3-29)	7.08-8.35 (7.49)	8.84-16.2 (10.8)	0.18-0.30 (0.26)	6.31	2.36	209.42	4.94
Italy*	4-23 (4-21)	3-30 (3-21)	3-31 (3-28)	114-115 (108)	203-399 (213+)	24.8-42.1 (29.3+)	12.18	13.76+	3526.7+	485.3+
Ger- many*	4-4 (4-7)	3-31 (3-27)	4-14 (4-15)	50.7-69.5 (70.8)	144-309 (167+)	6.86-14.0 (7.0+)	9.62	4.19+	2014.6+	84.35+
Spain*	4-22 (4-27)	4-6 (3-26)	4-7 (4-1)	114-129 (105)	274-495 (248+)	28.1-48.4 (25.6+)	10.44	10.32+	5316.9+	548.5+
USA*	5-12 (NA)	4-19 (4-9)	4-22 (4-16)	1400-1474 (943+)	1520-2410 (1204+)	90-140 (71.1+)	12.21	5.9+	3649.5+	215.3+
UK*	6-30# (NA)	4-20 (4-10)	4-20 (4-30)	11789# (165+)	249-392 (195+)	33-62 (29.4+)	14.21	15.1+	2932.2+	442.5+

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

Table 1. Summary of the predictions made on May 5, 2020. **Notes:** Countries with “*” sign after a country name indicates that the epidemic is still developing, although some of the particular event might have happened. The number in the parenthesis is the actual data. A “+” sign indicates the latest value in the particular case that is still developing and that particular event has not happened yet. The ‘#’ indicated the result might be questionable, for the lack of recovered case numbers.

IPM= Infection per Million of Population. FPM=Fatality per Million of population.

6. Discussions and Conclusion.

There are actually two modeling approaches: mechanistic and statistical models. Mechanistic model is based on precisely defined epidemical parameters. Without *a priori* knowledge of the disease characteristic and existing social contact conditions, however, many of the parameters used are based on assumptions and educated guesses. Nevertheless, they are useful for exploring different scenarios for the policy decision to fight an epidemic prior to the rapid rise of the outbreak. In this paper, we introduced a statistical model that could also serve an important role complementary to the mechanistic models, to be used after the outbreak started. It can provide detailed tracking and prediction of the course of the epidemic. For the statistical model to work well, good quality data are indispensable. Our study indicated that with quality data, tracking and prediction of critical events such as the peak and the turning point could be accurate within days.

Our model is supported by underlying theoretic foundation and validated by the existing data from the region where the pandemic is waning. Many parameters characterizing an epidemic can be determined from local-in-time data. Because it is based on general epidemiological principles we suggest that our approach could be applied not just to the current Covid-19 epidemic, but also generally to future epidemics.

Importantly, we made explicit the concept of “suppressed equilibrium” as an end state of an epidemic in addition to the traditional “herd immunity” state. Based on

869 the traditional mechanistic model, an epidemic ends after a high percentage of the
870 population is infected and recovered hence acquiring immunity. This is the so-
871 called “herd immunity” idea. For COVID-19, which we found to be very contagious,
872 more so than previously thought, the “herd immunity” end would require almost all
873 of the population be infected and therefore would bring unthinkable toll in the
874 number of people sick and dead. A second way for an epidemic to end is with strict
875 contact-reduction measures, so that although a large pool of susceptible population
876 still exists, the portion that an infected person comes in contact with is reduced by
877 the measures adopted, again leading to the Effective Reproductive Number less than
878 1. Unlike the first end state mentioned above, this “suppressed” state is
879 “parametrically unstable” in the sense that if the social distancing measures are
880 relaxed before the epidemic ends or new infection is imported after the first wave
881 ends, the epidemic will rebound, as a large portion of the population is still
882 susceptible. For this second state to be a stable equilibrium, the social-distancing
883 measures, and quarantine of cross-border visitors need to be maintained until it is
884 clear that the disease has died off. It is this second state that most countries are now
885 aiming for.

886
887 The second state will have a much earlier date for the new cases to peak and the
888 epidemic to end. For the US the peak of the newly infected cases is April 7-11. The
889 epidemic is estimated here to end in the first week of June, assuming the current
890 social distancing/stay-at-home measures are maintained till then, and after that
891 date import of infected from abroad is prevented by strict quarantine of visitors.
892 These assumptions now do not look like they would hold as states begin to open
893 their business prematurely. For the US, we predict that the total number of infected
894 cases is 1.5 to 2.4 million, dependent on the assumption. These are symptomatic
895 cases that require hospitalization.

896
897 Since it is the goal of most countries to eventually approach the “suppressed
898 equilibrium”, it is important to note that the slowing growth of the incidence (daily
899 newly infected) that is observed is not a function of biology, but is a result of
900 contact-reduction, which is social science. The mechanism of the exhaustion of
901 susceptibles is not relevant anymore as the number of infected is such a very small
902 percentage of the susceptible population. Therefore it is not necessary to use a
903 model, such as SIR or SEIR, to keep track of the change in susceptibles after the start
904 of the outbreak. If these mechanistic models are to be used to track the progress of
905 an outbreak, the infection rate a in the models needs to be adjusted constantly, but it
906 is not clear how to adjust it. In any case, the solution is responding to input of
907 changing model parameters, and not to the natural biological evolution built into the
908 model structure, such as recovery and immunity upon recovery.

909
910 This is a new way to look at the field of epidemiology. The observed slowing of the
911 growth of incidence and the cresting of the epidemic curve are dominantly the
912 result of a reduction of “infectivity”, as first proposed by John Brownlee one
913 hundred years ago, except that the change is not due to biology---- from the “loss of
914 the infecting power on the part of the organism” as he thought, at least not in the

915 case of COVID-19, but as the consequence of reduced contacts among people in
916 lowering the Reproduction Number, which is defined as the number of other people
917 one infected person would infect. This measure is a product of the average number
918 a typical infected person comes in contact with and the probability that a contacted
919 person becomes infected. While the study of the latter is biology and medicine, that
920 for the former is social science and public health.
921
922
923

Appendices: THEORY

924

925

926 **Definition:** Let $I(t)$ be the number of active infected at time t . Its change is given

927 by;

928
$$\frac{d}{dt}I = N(t) - R(t),$$

929 where $N(t)$ is the number of newly infected, and $R(t)$ that of the newly recovered or
930 removed (dead). The term: Acting Infected Case (*AIC*) number is used to denote the
931 confirmed $I(t)$ when we deal with data.

932

933 Let t_p , the turning point defined as the peak of the active infected number. At this

934 point maximum medical resource is needed. This maximum occurs when

935
$$\frac{d}{dt}I = 0, \text{ implying } N(t_p) = R(t_p).$$

936 There is no need to first find $I(t)$ to locate this peak. Figure 8 shows how this is
937 determined locally. This local-in-time metric avoids the accumulation of poor early
938 data. After the turning point, the newly recovered starts to exceed the newly
939 infected. The demand for medical resources, such as hospital beds, isolation wards
940 and respirators, starts to decrease.

941

942 We consider a solitary outbreak. Let $t = 0$ be when the first infection began. For
943 Wuhan, China, this date is near the end of 2019, perhaps even earlier. Let t_B be the
944 beginning of the better quality data. This time is beyond the initial incubation
945 period of the disease and it can be assumed that at that time there is already a
946 population of infected, some of them asymptomatic but nevertheless infectious.

947

948 Let $X(t,s)$ be the number of infected cases at time t , with s being the “age”
949 distribution, i.e. number of days sick.

950 The total number of infected is given by:

951
$$I(t) = \int_0^T X(t,s) ds .$$

952 After being sick for T days, a patient either recovers or is removed (dead). T is called
953 the recovery period (or removal period). It is also called the infectious period if the
954 patient is infectious during this period. Of course its value varies by patient and by
955 the efficacy of treatment for each hospital. For the removed it also depends on the
956 age of the patient and whether there are underlying medical conditions. Only a
957 mean recovery period is obtainable from data, and so this is in reality a statistical
958 quantity. We will discuss later how this statistical quantity can be obtained from
959 data.

960

961 **Conservation law** (see ref(22)):

962 After first infected and until removed or cured, we have:

$$dX(t,s) = \frac{\partial}{\partial t} X \cdot dt + \frac{\partial}{\partial s} X \cdot ds = 0, \quad 0 < s < T.$$

963 So, since $ds/dt = 1$,

$$\frac{\partial}{\partial t} X + \frac{\partial}{\partial s} X = 0.$$

964 This equation is to be solved using the method of characteristics as

$X(t,s) = \text{constant}$ along characteristics defined by $ds/dt = 1$.

Boundary condition: $X(t,0)$, specifies the "birth" process,

965 i.e. how the disease spawns newly infected (with "age" $s = 0$).

Initial condition: $X(0,s) = 0$ for $s > 0$, specifies the initial age distribution at $t = 0$

966 There are two types of characteristics:

967 (i) $s > t$, (ii) $s < t$.

968 The first type of characteristics intersects the $t = 0$ axis, and since the initial
969 condition is zero, we have the solution:
970

971 $X(t,s) \equiv 0$ for $s > t$.

972

973 That is, there is no infected population who is sick for more days than the lapsed
974 time since the first infection occurred.

975

976 For the second type of characteristics, $t > s$ the solution is
977

978 $X(t,s) = f(s-t)$

979 with the form of f to be determined by the boundary condition. Even without
980 determining the form of f we have the following general results:

981 For $t > T$, and therefore $t > s$:

$$\begin{aligned} 982 \quad I(t) &= \int_0^T X(t,s) ds = \int_0^T f(s-t) ds \\ &= \int_{t-T}^t f(p) dp. \end{aligned}$$

$$983 \quad \frac{d}{dt} I = f(t) - f(t-T).$$

984

985 Since the rate of increase of $I(t)$ is by definition equal to the newly infected number,
986 $N(t)$, minus the newly recovered (or removed) number, $R(t)$, we have:

$$987 \quad N(t) - R(t) = \frac{d}{dt} I = f(t) - f(t-T).$$

988

989 For a fatal disease with low fatality rate, where almost all infected cases eventually
990 recover after a hospital stay of T days, we can identify

991 $f(t)$ with $N(t)$, and $f(t-T)$ with $R(t)$.

992 If the disease has a non-negligible fatality rate, we include the dead in $R(t)$.

993 **Main Result:** The daily newly recovered/removed number $R(t)$, is related to the
994 daily newly infected number $N(t)$ as, for $t > T$:

995

996

997

$$R(t) = N(t - T).$$

998 The second type of characteristics intersects the boundary $s = 0$. The boundary
999 condition itself needs to be solved as a function of t to describe how new infection
1000 (at $s = 0$) occurs. This can be done using a birth model, such as Eq. (1.56) in (23).

1001

1002 *Boundary condition:* Following Murray's Eq. (1.56), the "birth" is assumed to be
1003 proportional to "parents" of suitable "age", with an age-dependent birth rate, $a(s)$.

1004

1005
$$X(t, 0) = \int_0^T a(s)X(t, s)ds .$$

1006 This equation needs to be solved numerically, except in the case of constant a . The
1007 solution for $X(t, 0) = g(t)$ is generally an increasing function of time.

1008 Since $X(t, s) = g(t - s)$, at some time $t_B > T$; $X(t_B, s) = g(t_B - s) \equiv f_0(s)$. That is, an
1009 "age" distribution can be converted into a time profile through the method of
1010 characteristics.

1011

1012 The solution of the integral equation is not presented here. For our purpose here it
1013 suffices to assume that the solution of this model yields a distribution with age that
1014 has a full spectrum $0 < s < T$ of infectives at a time t_B , long after a full incubation
1015 period has passed.

1016
$$X(t_B, s) = f_0(s) = A \exp\left\{-\frac{(s - s_0)^2}{2b^2}\right\}; A \text{ independent of } s.$$

$$b = \frac{1}{2}T.$$

1017 This distribution is justified as follows:

1018 To the right of the peak in "age" distribution, it is easy to understand that the
1019 numbers for those who are "older" should be less because they were infected earlier
1020 during an earlier stage of the epidemic. As the epidemic grows exponentially, there
1021 are more and more "younger" infectives. To the left of the peak in "age" distribution
1022 many are at various stages of incubation. Those who are newly infected may be
1023 large in number, but they are less infectious and contribute less to the growth of the
1024 subsequent infection, and since they have not developed symptoms, they are less
1025 likely to be hospitalized, tested and contribute to the "case" record. For Covid-19,
1026 the peak infectiousness occurs in a period just before and after the onset of
1027 symptoms, around 5-6 days since first infected (23). The above discussion gives an
1028 epidemiological justification for a Gaussian-like distribution of the "age" distribution
1029 of the "new cases", more appropriately called "new hospitalization per day".

1030

1031 It is important to point out that the assumption of less “cases” to the left of the peak
 1032 is the only “filter” that serves to limit our theoretical discussion to the data of
 1033 symptomatic cases. If all infected were included, $f_0(s)$ should have a maximum at
 1034 $s = 0$.

1035 Therefore the solution is, for $t > t_B > 0$:

$$1036 \quad X(t,s) = X(t_B, s-t) = f_0(s-t) \\ = A \exp\left\{-\frac{(s-s_0-t)^2}{2b^2}\right\}.$$

1037

$$1038 \quad N(t) = A \exp\left\{-\frac{(t-s_0)^2}{2b^2}\right\} = A \exp\left\{-\frac{(t-t_N)^2}{2b^2}\right\}$$

$$R(t) = N(t-T) = A \exp\left\{-\frac{(t-t_R)^2}{2b^2}\right\},$$

1039 where t_N is the peak of $N(t)$, and $t_R = t_N + T$ is the peak of $R(t)$.

1040 Both distributions are Gaussians.

1041

1042 For $t_B < t < T$,

$$1043 \quad I(t) = \int_0^t X(t,s) ds + \int_t^T X(t,s) ds \\ = \int_0^t f(s-t) ds + \int_t^T 0 ds \\ = \int_{-t}^0 f(p) dp$$

$$1044 \quad \frac{d}{dt} I = f(-t) = A \exp\left[-\frac{(t-s_0)^2}{T^2}\right].$$

$$1045 \quad N(t) = A \exp\left[-\frac{(t-t_N)^2}{T^2}\right]$$

$$R(t) = 0.$$

1046 Again, $N(t)$ is Gaussian, but there is no recovered or removed during this early stage.

1047

1048 **Main Result:** The natural logarithm of the ratio of N and R is a linear function of
 1049 time for $t > T$:

1050

$$1051 \quad NR(t) = \frac{N(t)}{R(t)} = \frac{N(t)}{N(t-T)} \\ = \exp\left\{-\frac{(t-t_N)^2}{2b^2} + \frac{(t-t_N-T)^2}{2b^2}\right\} = \exp\left\{\frac{T^2}{2b^2} - \frac{T(t-t_N)}{b^2}\right\}.$$

$$\log NR = -\frac{T(t-t_N) - \frac{1}{2}T^2}{b^2}$$

1052 a linear function of t .

1053

1054 **Heterogeneous Data**

1055 The above results are obtained for the case of a single introduction into a region of
 1056 infected at $t=0$ and we solve for the subsequent development of the epidemic from
 1057 that single source. Consider now a large region consisting of a number of small
 1058 regions, and the “seeding” of the infected occurs at different times for different
 1059 regions. The large region could be China, and the first infection could be Wuhan,
 1060 Hubei and then the regions outside Hubei. Then we may have for the China as a
 1061 whole data for the newly infected a sum of several Gaussians staggered in time. As
 1062 long as the Gaussians are not separated so much that there are different peaks in the
 1063 combined data, the combined data can still be considered as Gaussian, as is the case
 1064 in the real data. However, the standard deviation σ of the combined Gaussian is
 1065 inevitably larger and is no longer given by b :

$$1066 \quad N(t) = \frac{B}{\sqrt{2\pi}\sigma_N} \exp\left\{-\frac{(t-t_N)^2}{2\sigma_N^2}\right\}.$$

1067 We still have $R(t) = N(t-T)$ since this result holds for each sub-region. The result
 1068 that $\log NR$ is a linear function of time still holds:

$$1069 \quad \log NR(t) = \log \frac{N(t)}{N(t-T)} = -\frac{T(t-t_N) - \frac{1}{2}T^2}{\sigma_N^2}.$$

1070 The slope of the straight line is T/σ_N^2 .

1071

1072 Since the hospital stay can act as a smoothing filter on $N(t)$ to yield $R(t)$, the
 1073 standard deviation for $R(t)$ could be slightly wider than that for $N(t)$. So we could
 1074 have two different Gaussians (but their integral over all time should be the same):

$$1075 \quad N(t) = \frac{B}{\sqrt{2\pi}\sigma_N} \exp\left\{-\frac{(t-t_N)^2}{2\sigma_N^2}\right\}; \quad R(t) = \frac{B}{\sqrt{2\pi}\sigma_R} \exp\left\{-\frac{(t-t_R)^2}{2\sigma_R^2}\right\}.$$

1076 Taking this into account, we have, denoting $T = t_R - t_N$:

$$1077 \quad \log NR - \log \frac{\sigma_R}{\sigma_N} = -\left\{\frac{(t-t_N)^2}{2\sigma_N^2} - \frac{(t-t_N-T)^2}{2\sigma_R^2}\right\} = -\frac{(\sigma_R^2 - \sigma_N^2)(t-t_N)^2 + 2\sigma_N^2 T(t-t_N) - \sigma_N^2 T^2}{2\sigma_N^2 \sigma_R^2}$$

1078

1079 As the values of σ_N and σ_R are very close based on the empirical data, the quadratic
 1080 term is always small comparing to the other terms for the length of time we are
 1081 considering here. Hence,

$$1082 \quad \log NR(t) = \frac{1}{\sigma_R^2} \left\{-T(t-t_N) + \frac{1}{2}T^2\right\},$$

1083 a linear function of time. Its slope is $-\frac{T}{\sigma_R^2}$.

1084 A caveat: For widely separated countries, such as China and US, the two Gaussian
 1085 peaks are separated in time. These two regions should be treated separately so that

1086 the data for each region has only a single peak. Similarly for Europe and China. See
1087 Appendix. There is also a problem with the aggregate data for the US consisting of
1088 different states with outbreaks separated in time.

1089
1090 **Result:** The natural logarithm of the ratio of two Gaussians of slightly different
1091 standard deviation is approximately a straight line, providing the data used is
1092 homogeneous temporally and spatially.

1093
1094 Sometimes, heterogeneity in one variable could be relaxed in an overall mean. The
1095 case in China is an example. The case could be treated as whole China even though
1096 China included Hubei and Wuhan and the very different region of China outside
1097 Hubei, for the outbreak occurred at the same time. But the cases in China and
1098 European countries and US could not be treated as a whole, for they are separated
1099 both spatially and temporally. Consequently, it would be impossible to talk about a
1100 single distribution for the global prediction.

1101
1102 **Time-varying σ and t_N**

1103 The above derivations continue to hold even for σ being a function of time. This
1104 becomes relevant when, for example, after the epidemic passes its peak the policy
1105 makers decide to relax the restrictions on social distancing. Such actions alter the
1106 course of the epidemic and create a longer tail of new cases. $N(t)$ is no longer
1107 Gaussian because it does not have fore-aft symmetry. Since such government
1108 actions are not included in the initial prediction, the prediction scheme needs to
1109 adapt using real-time data after the peak. Our method allows for it.

1110
1111
1112 **The peak infected cases:**

1113 Writing $N(t_N) = N(t_B) \exp\{\int_{t_B}^{t_N} n(t) dt\}$, and noting $n(t) = \frac{d}{dt} \log N(t) = -\frac{(t-t_N)}{\sigma_N^2}$, the

1114 exponent is $\int_{t_B}^{t_N} n(t) dt = \frac{(t_N - t_B)^2}{2\sigma_N^2} = \frac{1}{2} n(t_B)(t_N - t_B)$. Hence the peak infected case number

1115 can be predicted, using the predicted value for t_N starting from a conveniently chosen
1116 time t_B , such as the latest time with data available, as:

$$1117 \quad N(t_N) = N(t_B) \exp\left\{\frac{n(t_B)(t_N - t_B)}{2}\right\} .$$

1118
1119 **Accumulated quantities.**

1120 To calculate $I(t)$ using reported data, only confirmed cases are used (We call it AIC).
1121 It is given by the accumulated newly confirmed cases minus the accumulated
1122 confirmed recovered. Since the accumulation of early poor data can introduce
1123 errors a more local-in-time formula is given as:

1124

1125
$$I(t) = \int_{-\infty}^t N(t)dt - \int_{-\infty}^t R(t)dt = \int_{-\infty}^t N(t)dt - \int_{-\infty}^t N(t-T)dt$$
$$= \int_{t-T}^t N(t)dt.$$

1126 That is, to find I at time t , one only needs to add up the daily newly infected case
1127 numbers for a period of T preceding t . This is an almost local-in-time property even
1128 for this accumulated quantity. For validation, we estimate the peak of the I case
1129 number on 18 February by computing the sum of daily newly infected case numbers
1130 for 15 days, from February 4 to February 18, which yields a peak value for the total
1131 infected cases on 18 February of 54,747. This is within 10% of the reported number
1132 of 57, 805, even after taking into account the deaths (by subtracting the
1133 accumulated deaths of 2,004 from our estimate).

1134
1135 The Total Infected Cases (TIC) is one of the most reported statistics:

1136
$$TIC(t) = \int_0^t N(t)dt.$$

1137 The total infected cases for the epidemic for a region after it is over is given by
1138 approximately:

1139
$$TIC_{\infty} = 2 \cdot TIC(t_N),$$

1140 assuming that $N(t)$ is approximately symmetric about its peak at t_N . In reality $N(t)$
1141 may not be symmetric and likely has a long tail. However, since the number of cases
1142 along the tail is small, the above approximation for the total is still good. Since TIC is
1143 officially available at any time t we will use that reported number, projected forward
1144 to the peak t_N , and then doubling it. Caveat: the official number may include
1145 accumulation of early bad data. Hopefully it is a small percentage of the sheer size of
1146 the TIC.

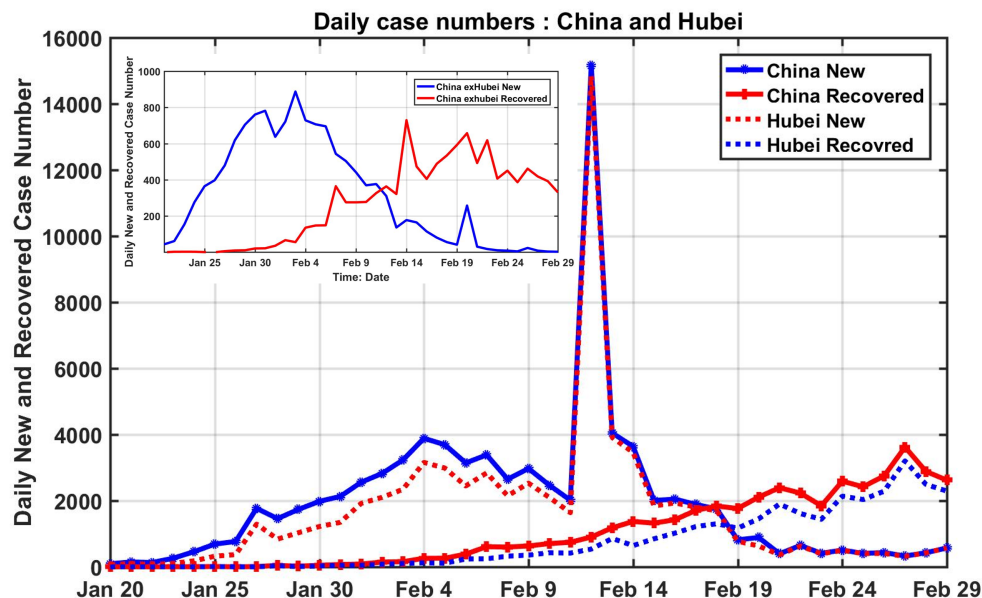
1147
1148 The peak AIC number can be predicted as

1149
$$I(t_p) = I(t_B) \exp\{\frac{1}{2} \alpha(t_B)(t_B - t_p)\},$$
 where t_B is the last available data before the

1150 turning point. It is assumed that $\alpha(t)$ lies on a straight line between t_B and t_p .

1151 This is the maximum load that health services need to plan for, and adopt “flatten the
1152 curve” policies to keep this number under the maximum resources available, such as
1153 hospital beds and isolation wards.

1154



1155
1156

Figure 8. The daily newly infected (in blue) and the daily newly recovered (in red), as a function of time for China as a whole (in solid lines) and Hubei (in dotted red and blue lines). The turning point is determined by when the red and blue curves cross.

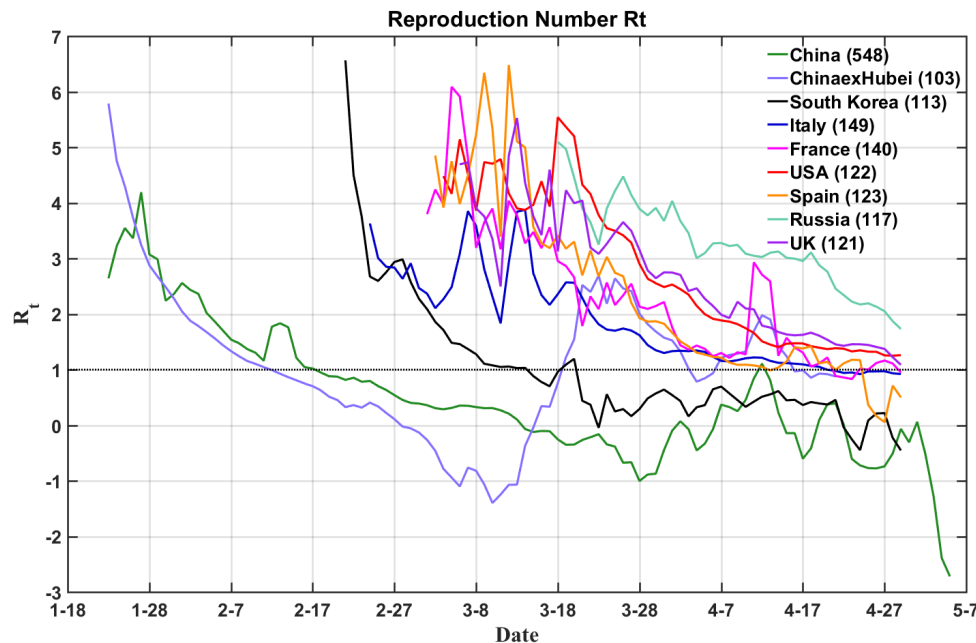
1159
1160
1161 **Inset:** For China outside Hubei.
1162

1163 1164 **The problem with data quality**

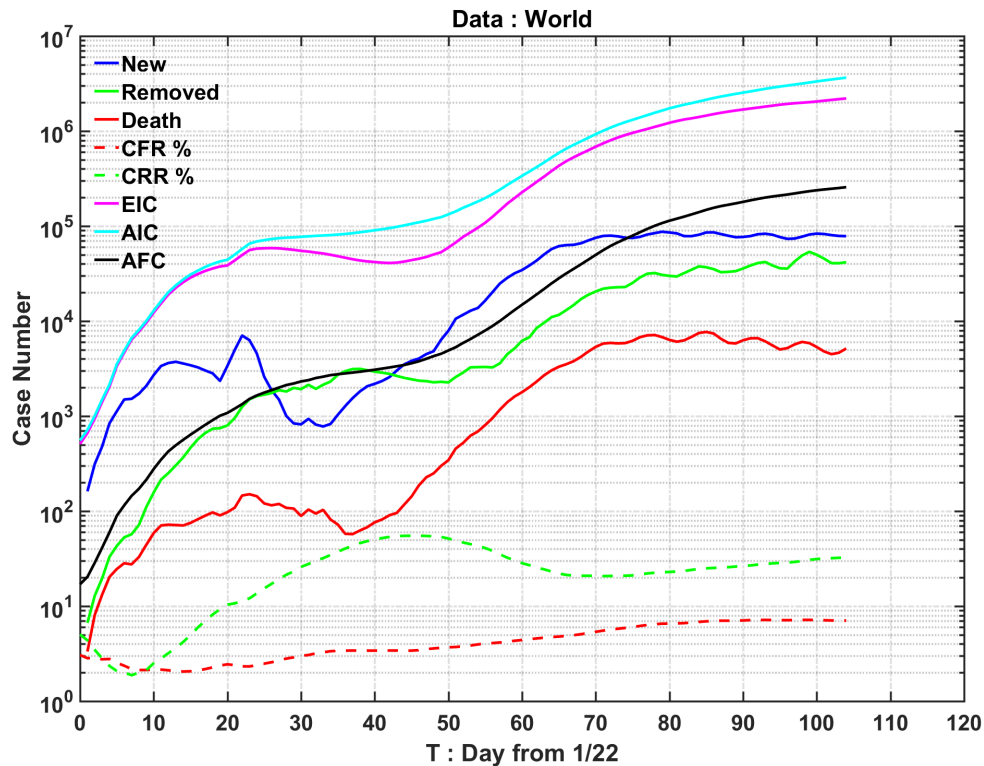
1165
1166 The performance of all data-driven models depends on the quality of data. In general,
1167 there are problems with data from all countries. Underreporting is universal to all
1168 countries at least during the initial period of the epidemic. The causes might be different.
1169 In China, for example, the data collection was erratic in the early days for lack of
1170 awareness of the novel virus. A more common cause was the shortage of test kits to
1171 identify the cases. The situation is made more complicate by the existence of
1172 asymptomatic cases. However, whether to include those cases or not depends on the
1173 definition. In this study, by definition, the cases are limited to confirmed ones with
1174 symptoms, for only those cases would tax the medical resources. In fact, the best any
1175 country can do is to record the confirmed new case accurately. The cases for death seem
1176 to be simple, for the number of deaths has attracted great attention by the government and
1177 public alike. Many models are designed to follow only death cases. However, even here,
1178 the numbers are not problem-free. For lack of testing, large number of unattributed
1179 COVID-19 deaths exists in every country. Finally, the most critical problem is for the
1180 recovered case numbers. It is important to have this record, for it is essential for the
1181 conservation law discussed above. Furthermore, the new to recovery/removal case ratio
1182 (NR Ratio) provides a robust prediction for the turning point of the epidemic.
1183 Unfortunately, the recovered case numbers had attracted the least attention. As a result, it
1184 is not recorded or reported in some countries, such as UK. Under this condition, many of

1185 the tracking and prediction can still be made as demonstrated in the main text. However,
1186 we strongly urge data for the recovery case to be recorded accurately, for it could not
1187 only provide accurate prediction for medical resource preparation, but also as a measure
1188 of the medical system efficacy.
1189

1190 Finally, a few words on data homogeneity. All models, statistical or mechanistic, require
1191 homogeneous condition. There are actually two types of homogeneity: temporal and
1192 spatial. We believe the temporal homogeneity is more critical. For example, China can
1193 be treated as a whole entity even with part of China, such Hubei and Wuhan under strict
1194 lockdown, and part of China is not. The result still makes sense because of its temporal
1195 homogeneity. Temporal inhomogeneity, however, would render the data nonsensical.
1196 Let us take the global condition as an example. There is a gap in time between the
1197 Chinese cases and the rest of the world (see Figure 9). No model should treat the two
1198 events as one. See Figure 10. On the other hand, the European country and the US could
1199 be treated as the World exChina, for the outbreak in European countries and the US
1200 happened at about the same time, even though they are spatially separated widely.
1201



1202
1203 **Figure 9.** This figure shows the Reproduction Number in real time starting on 20
1204 January. In this presentation, one can see the pandemic is consisted of a serial
1205 event with severe data inhomogeneity in both space and time. The outbreaks were
1206 almost over in China and South Korea before the earliest cases in Italy and Spain
1207 started to rise. There were 40 days difference between China and South Korea on
1208 one hand, and the cluster of the pandemic flared up in European countries and the
1209 US on the other hand.
1210
1211



1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237

Figure 10. China and the rest of the world should be treat separately.

Acknowledgements: NEH and FQ are supported by the National Natural Science Foundation of China under Grant 41821004. KKT's research is supported by the Frederic and Julia Wan Endowed Professorship.

Competing Interests: The authors declare no competing interests.

Data Availability: All data in this study are publicly available from World Health Organization (WHO) at <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>

and on the Daily Brief site of the China's National Health Commission at <http://en.nhc.gov.cn/>

The Korean data is available at

<https://sa.sogou.com/new-weball/page/sgs/epidemic>

Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE

<https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6>

[https://github.com/CSSEGISandData/COVID-](https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series)

[19/tree/master/csse_covid_19_data/csse_covid_19_time_series](https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series)

1238 **References**

- 1239 1. Cyraboski D. When will the coronavirus outbreak peak? Nature News. 2020.
1240 2. Adams D. Modelers struggle to predict the future of the COVID-19 pandemic.
1241 The Scientist. 2020.
1242 3. Kermack W, McKendrick A. A contribution to the mathematical theory of
1243 epidemics. . Proc Roy Soc London 1927;A115:700-21.
1244 4. Ferguson NM. Impact of non-pharmaceutical interventions (NPIs) to reduce
1245 COVID-19 mortality and healthcare demand. *Spiral*. 2020.
1246 5. Flaxman S. Report 13: Estimating the number of infections and the impact of
1247 non-pharmaceutical interventions on COVID-19 in 11 European countries. *Spiral*.
1248 2020.
1249 6. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High
1250 contagiousness and rapid spread of severe acute respiratory syndrom coronavirus 2.
1251 *Emerging Infectious Disease*. 2020;26(7).
1252 7. Bendavid E, Mulaney B, Sood N, Shah S, Ling E, Bromley-Dulfano R, et al.
1253 COVID-19 antibody seroprevalence in Santa Clara county, California. *medRxiv*. 2020.
1254 8. Susser M, A. A. Causes of death: epidemic, infectious, and zymoptic diseases.
1255 *Vital Statistics: A memorial volume of selections from the reports and writings of*
1256 *William Farr*: Scarecrow Press Inc.; 1975. p. 317-21.
1257 9. Murray C. Forecasting COVID-19 impact on hospital bed-days, ICU-days,
1258 ventilator-days and deaths by US state in the next 4 months 2020.
1259 10. Jewell NP, Lewnard JA, Jewell BL. Caution Warranted: Using the Institute for
1260 Health Metrics and
1261 Evaluation Model for Predicting the Course of the COVID-19 Pandemic. *Annals of*
1262 *Internal Medicine*. 2020;173(3).
1263 11. Woody S, Tec M, Dahan M, Gaither K, Lachmann M, Fox SJ, et al. Projections
1264 for first-wave COVID-19 deaths
1265 across the US using social-distancing
1266 measures derived from mobile phones. Preprint. 2020.
1267 12. Fine PEM. John Brownlee and the measurement of infectiousness: An
1268 historical study in epidemic theory. *Jornal of Royal Statistical Society A*.
1269 1979;142:347-62.
1270 13. Brownlee J. Statistical studies in immunity. The theory of an epidemic. *Proc*
1271 *Roy Soc Edinburgh*. 1907;26:484-521.
1272 14. Brownlee J. Statistical studies in immunity:the theory of an epidemic. *Proc R*
1273 *Soc Edinburg*. 1907;26:484-521.
1274 15. Huang NE, Qiao F. A data driven time-dependent transmission rate for
1275 tracking an epidemic: a case study of 2019-nCoV. 65, 425-427, . *Sci Bull*.
1276 2020;65:425-7.
1277 16. Delamater PL, Street EJ, Leslie TF, Yang YT, H. JK. Complexity of the Basic
1278 Reproduction Number (R0). *Emerging Infectious Diseases*. 2019;25.
1279 17. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential
1280 domestic and international spread of the 2019-nCoV outbreak originating in Wuhan,
1281 China: a modelling study. . *Lancet*. 2020;395(10225):689-97.

- 1282 18. Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD. How will
1283 country-based mitigation measures influence the course of the COVID-19 epidemic?
1284 Lancet. 2020;395:931-4.
- 1285 19. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory
1286 syndrome reveal similar impacts of control measures. Ameri J Epidemiology.
1287 2004;160:509-16.
- 1288 20. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD.
1289 Pandemic potential of a strain of influenza A (H1N1): Early findings. Science.
1290 2009;324:1557-61.
- 1291 21. Thakkar N, Burstein R, Hu H, Selvaraj P, Klein D. Social distancing and
1292 mobility reductions have reduced COVID-19 transmission in King County, WA.
1293 Institute for Disease Modeling; 2020.
- 1294 22. Murray JD. Mathematical Biology I. Third ed: Springer. 551 p.
- 1295 23. Lauer SA, Grantz KH, Bi Q, Jones FK. The incubation period of coronavirus
1296 disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and
1297 application. Annals of Internal Medicine. 2020;172:577-82.
- 1298