

# On Identifying and Mitigating Bias in the Estimation of the COVID-19 Case Fatality Rate

Anastasios Nikolas Angelopoulos<sup>†</sup>, Reese Pathak<sup>†</sup>, Rohit Varma<sup>◇</sup>, and Michael I. Jordan<sup>†, ‡</sup>

<sup>◇</sup> Southern California Eye Institute, CHA Hollywood Presbyterian Medical Center, Los Angeles

<sup>†</sup> Department of Electrical Engineering and Computer Science, UC Berkeley

<sup>‡</sup> Department of Statistics, UC Berkeley

{angelopoulos, pathakr, jordan}@cs.berkeley.edu, rvarma@sceyes.org

## Abstract

The relative case fatality rates (CFRs) between groups and countries are key measures of relative risk that guide policy decisions regarding scarce medical resource allocation during the ongoing COVID-19 pandemic. In the middle of an active outbreak when surveillance data is the primary source of information, estimating these quantities involves compensating for competing biases in time series of deaths, cases and recoveries. These include time- and severity-dependent reporting of cases as well as time lags in observed patient outcomes. In the context of COVID-19 CFR estimation, we survey such biases and their potential significances. Further, we analyze theoretically the effect of certain biases, like preferential reporting of fatal cases, on naive estimators of CFR. We provide a partially corrected estimator of these naive estimates that accounts for time lag and imperfect reporting of deaths and recoveries. We show that collection of randomized data by testing the contacts of infectious individuals regardless of the presence of symptoms would mitigate bias by limiting the covariance between diagnosis and death. Our analysis is supplemented by theoretical and numerical results and a simple and fast [open-source codebase](#).

## 1 Introduction

As of May 18, 2020, the 2019 novel Coronavirus (SARS-CoV-2) outbreak has claimed at least 317,000 lives out of 4.8 million confirmed cases worldwide, of which 1.8 million recovered ([Dong et al., 2020](#)). Because the basic reproduction number  $R_0$  of the virus is high (estimated to fall between 2 and 3 by [Liu et al. 2020](#)), public health organizations and local, state, and national governments must allocate scarce resources to populations especially susceptible to death during this pandemic. Therefore it is critical to have good estimates of the proportion of fatal infections of COVID-19: this quantity is referred to as the absolute case fatality rate (CFR)<sup>1</sup>. It is additionally important to understand the relative CFRs between different sub-populations (i.e., the ratio of their absolute CFRs). We view the relative CFR as a useful target for data-informed resource allocation protocols because it is a key measure of relative risk. Indeed, the absolute CFR is a measure of absolute severity only for a particular population, since it averages out effects of medical care, age, geography, genetics, and more. Practical decisions will ultimately be made based on coarse

---

<sup>1</sup>An important caveat at the outset: As Box 1 of [Lipsitch et al. \(2015\)](#) points out, “The CFR... is an ambiguous term, as its definition and value depend on what qualifies an individual to be a ‘case.’” In this paper, we are defining the CFR as the proportion of fatal infections; i.e., the proportion of deaths among all COVID-19 infected individuals. This is a version of the CFR that is often called the infection fatality rate (IFR). As we will discuss, even this definition is ambiguous for many reasons, including the cause of death. While no perfect definition exists, due in part to the many biases described in Section 2, in any given analysis it will be important to choose a pertinent definition, and to take care when making comparisons between analyses that have chosen different definitions.

stratifications of these covariates; for example, a relative CFR that is specific to a geographical region may be needed for resource planning and allocation. Similarly, a relative CFR that is specific to sex or race is often sought to monitor for and ensure equitable treatment of patients within hospitals across demographics. To facilitate such planning, we target the relative number of deaths among total cases between groups of people (*e.g.*, senior citizens in Italy) as a critical measure of relative risk that informs decisions affecting human lives. Other such measures include prevalence and risk of hospitalization.

It is widely believed that the naïve estimator of CFR,  $E_{\text{naive}}$ , obtained from a simple ratio of reported deaths to reported cases (and which has a value of 6.6% when applied to the data of May 18, 2020), is biased (Fauci et al., 2020; Battegay et al., 2020). Indeed, an extensive epidemiological literature has asserted this bias and presented methods that attempt to mitigate it (Donnelly et al., 2003). Bias-mitigation methods are also present in a large literature on survey sampling and weighting (Gelman, 2007). Despite this academic background, naïve estimates continue to be used, reported, and cited in major publications (The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team, 2020; Z. Wu & McGoogan, 2020; Lipsitch, 2020).

Since publicly available health surveillance data for COVID-19 are heterogeneous and partially observed, it is problematic to assert that any estimator uniformly outperforms the naïve estimator. A variety of competing (and unknown) biases, both negative and positive, could conceivably cancel, causing the naïve estimator to be closer to the true CFR despite its theoretical inadequacy. Statistical wisdom would suggest that the conundrum of conflicting biases can be remedied by studying the multi-stage process that relates data obtained by surveillance sampling to the populations that are the target of inferential assertions. It is the goal of this paper to present such a statistical perspective and explore some of its consequences for COVID-19.

Clarity on the potential biases underlying the use of data from surveillance sampling can help to determine what additional datasets may be needed to mitigate bias. Examples which will inform our discussion include the New York seroprevalence study reported in Goodman & Rothfeld (2020), which helped correct significant under-ascertainment of mild cases, and Verity et al. (2020), who made use of individualized case data, polymerase-chain reaction (PCR) prevalence data, and Bayesian inferential methods, resulting in a CFR estimate of 1.38%. These studies can improve public-health response to COVID-19 insofar as they are accompanied by an understanding of their implicit assumptions, including putative control of possible biases.

The remainder of this paper is organized as follows. In Section 2, we provide a statistical perspective on the many potential biases affecting absolute and relative CFR estimation. In Section 3, we employ the general perspective to isolate some restricted contexts in which two naïve estimators are unbiased, with implications on the need for contact tracing. In Section 4, we consider how model-based inference can expand the contexts in which unbiased estimation is possible. We provide an illustrative example, showing how an (approximate) maximum likelihood estimator from Reich et al. (2012) can be applied to correct bias from relative reporting rates of fatal and resolved cases using only surveillance data and an approximate horizon distribution of deaths. We discuss how the general principle of coping with incomplete data via Poisson approximation and a log-linear likelihood model can be more widely applied. In Section 5, we present results of this method on COVID-19 data. Finally, in Section 6, we give a mathematical justification for contact tracing as a data-collection methodology (Eames & Keeling, 2003; Chowell et al., 2009), and discuss how it would mitigate many of the problematic biases at their source.

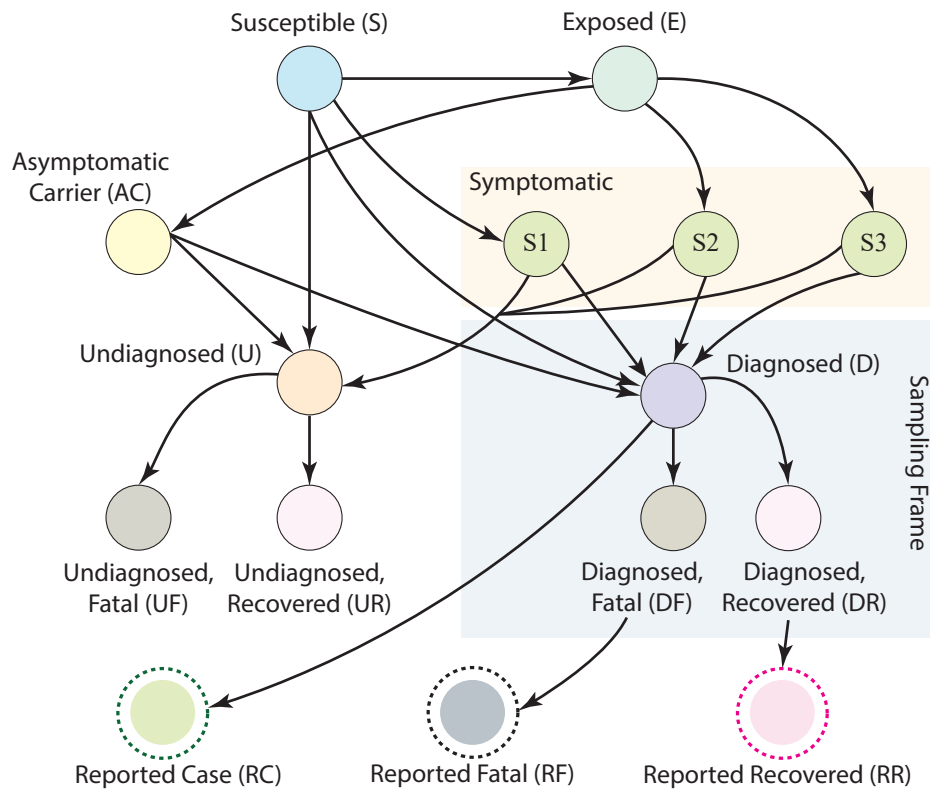
## 2 Sources of bias in COVID-19 surveillance data

In Figure 1, we present a graphical model that captures aspects of the data-generating process for COVID-19 surveillance data. Graphical models provide a general formal language for reasoning about the probabilistic and causal structure of collections of random variables (see, e.g., [Jordan 2004](#)). In this paper, it suffices to think of the model in Figure 1 informally as a depiction of dependencies among population-level and sampled quantities in surveillance data. Our objective is to consider biases that may arise along each edge of the graph. Prior work on SARS, H1N1, H7N9, H5N1, MERS, and HIV has identified or even quantified many of these biases ([Atkins et al., 2015](#); [Woodruff et al., 2014](#)); see [Lipsitch et al. \(2015\)](#) for a review. The diagram also depicts the population eligible to be sampled (the “sampling frame”) through collection of surveillance data from standard hospital reports and death certificates. Asymptomatic carriers are excluded from the sampling frame because testing is currently not recommended or available for asymptomatic individuals. We roughly categorize biases as: under-ascertainment of mild cases, time lags, interventions, group characteristics, and imperfect reporting and attribution. Extensive (but not comprehensive) discussion of the magnitude and direction of these biases corroborated by COVID-19-specific evidence is included in the subsections below.

These competing biases can often be expressed in terms of ratios of edge weights (conditional probabilities) from Figure 1. For example, with  $A \rightarrow B$  denoting the value of the edge between nodes A and B, if the probability of reporting a fatal case is greater than reporting an infection ( $\frac{DF \rightarrow RF}{D \rightarrow RC} > 1$ ), then ignoring other biases,  $E_{\text{naive}}$  will be upwardly biased by a factor  $b > 1$ . However, because biases can compete with one another, this does not mean  $\frac{1}{b}E_{\text{naive}}$  is always a better estimator than  $E_{\text{naive}}$ . Speaking loosely, the bias incurred by under-ascertainment of asymptomatic cases could be  $\frac{1}{b}$ , canceling out the former bias. Accordingly, the total error of any estimator is indeterminate. Therefore, the use of estimators based on surveillance data requires being clear on underlying assumptions. Statisticians and public health officials should proceed with multiple estimation strategies, armed with an understanding of the accompanying biases, and they should endeavor to collect additional data that mitigates the biases (as we describe in Section 6).

Figure 1 conveys two important takeaways regarding the estimation of CFR. First and foremost, information about edges outside the sampling frame cannot be inferred from data within the sampling frame alone. Even within the frame, data is compromised by the biases listed in the subsections below. Each incoming edge to  $D$  can bias CFR estimates up or down, depending on its ratio to other incoming edges. This cannot be disentangled only by looking at the value of  $D$ . Secondly, even within the sampling frame, the relative values and time lags of  $D \rightarrow DF$  and  $D \rightarrow DR$  affect the estimation of CFR. However, these edges may be the only ones subject to correction using population-level surveillance data alone. This motivates our choice of an illustrative estimator of relative CFR, adapted from [Reich et al. \(2012\)](#). In particular, the estimator is based on assumptions under which estimation of relative CFR is possible while correcting for the relative values and time lags of  $D \rightarrow DF$  and  $D \rightarrow DR$ .

In the following subsections, we will cite evidence for the existence, magnitudes, and directions of certain biases from Figure 1. Our analysis was done in mid-April. We categorize biases resulting from one of five phenomena: under-ascertainment of mild cases; time lags; interventions; group characteristics; and imperfect reporting and attribution.



Edge	Notes and some potential sources of sampling and estimation bias
$S \rightarrow E$	Social distancing, occupation, family size, behavior.
$S \rightarrow U$	COVID-negative people who are not diagnosed.
$S \rightarrow D$	Assay specificity and sensitivity, test availability.
$S \rightarrow S1$	Group characteristics such as genetics and immunity. Exposure to flu.
$E \rightarrow AC, S2$	Infectious dose, route of transmission, and group characteristics.
$E \rightarrow S3$	Presence of other underlying medical conditions.
$AC \rightarrow U, D$	Random sampling, contact tracing, test availability.
$S1 \rightarrow U$	Assay specificity and sensitivity, case severity.
$S2 \rightarrow U$	Assay specificity and sensitivity, case severity, test availability.
$S3 \rightarrow U$	Misattribution of symptoms, assay specificity and sensitivity, comorbidities.
$S1 \rightarrow D$	Misattribution of symptoms, assay specificity and sensitivity, group characteristics such as access to and effectiveness of medical care.
$S2 \rightarrow D$	Delays in seeking care, government intervention, contact tracing, group characteristics.
$S3 \rightarrow D$	Assay specificity and sensitivity, contact tracing and test availability.
$U \rightarrow UF, UR$	Group characteristics, particularly comorbidities.
$D \rightarrow DF$	Misattribution, group characteristics, imperfect reporting.
$D \rightarrow DR$	Survey non-response, group characteristics, imperfect reporting.
$D \rightarrow RC$	Imperfect reporting: errors, change in case definition, release of incorrect data. Time lag.
$DF \rightarrow RF$	Imperfect reporting, national reporting guidelines, ease of reporting.
$DR \rightarrow RR$	Imperfect reporting, national reporting guidelines, ease of reporting.

**Figure 1. Some sources of bias arising from COVID-19 health surveillance data.** For simplicity, nodes in this diagram can be thought of as representing the number of people in the labeled state. Each edge represents a conditional probability of transitioning between states and has an associated time lag. S1 represents COVID-19 negative people with flu symptoms. S2 represents people with symptoms *caused* by COVID-19. S3 represents COVID-19 positive people with symptoms caused by another underlying health condition. The sampling frame of population surveillance data based on standard hospital reporting is in light blue; values of edges outside the sampling frame cannot be inferred only with data from within the sampling frame. Relative time lag introduces bias across all edges.

## 2.1 Under-ascertainment of mild cases

Diagnosing severe cases more often than mild cases will falsely increase CFR. In Figure 1, this bias corresponds most directly to spuriously increasing  $AC \rightarrow U$  and  $S2 \rightarrow U$  and/or decreasing  $AC \rightarrow D$  and  $S2 \rightarrow U$ . This bias may have high magnitude, since the true number of infections is likely to be several times as high as the reported number of cases in countries where testing is limited (Fauci et al., 2020). The significantly lower CFR in South Korea, a country with widespread testing, corroborates this explanation (Dong et al., 2020). A recent serology study from New York City (reported in Goodman & Rothfeld 2020) suggests the prevalence of COVID-19 is 21.2%, much higher than the confirmed case count. This indicates that the number of infections may be much larger than surveillance data implies globally.

## 2.2 Time lags

Deaths and recoveries are reported after cases are confirmed, which artificially decreases the naïve CFR (Wilson et al., 2020). More specifically, when a number of new cases is reported without delay,  $E_{\text{naïve}}$  becomes biased downward since the deaths yet to occur from the new cases will be missing from the numerator. In Figure 1, this means a time lag is incurred across edges  $D \rightarrow DF$  and  $D \rightarrow DR$ . The value of this time lag depends on how early on in the disease’s course a patient is diagnosed. The median value of the lag across  $D \rightarrow DF$  was estimated by Linton et al. in China in early February to be 6.7 days (Lognormal 95% CI: 5.3, 8.3) (Linton et al., 2020). Tracking individual cases and including them only after death or recovery would solve this problem. However, data on individuals is rare, as hospitals and/or governments generally report population-level data only.

Fortunately, these edges are within the sampling frame, so we can have some hope of correcting the bias incurred by this time lag. In Section 4, we implement an estimator that handles this correction directly, and discuss the many further assumptions that must be made to assure its validity even within the sampling frame. In reality however, time lags affect every edge in Figure 1. The “incubation period,” for example, creates lag across edges  $E \rightarrow AC, S1, S2, S3$ , and was estimated at median 4.3 days based on Linton et al. The time between onset and hospital admission, which is not perfectly represented by the graphical model, creates lag across edges  $S1, S2, S3 \rightarrow D$  and had an estimated median of 1.5 days among recovered cases and 5.1 days among fatal ones (Linton et al., 2020). The large discrepancy between hospitalization lags of patients with different outcomes suggests the presence of an unknown bias factor. For example, earlier hospitalization may result in more effective treatment, canceling out the propensity of severe cases to seek care more quickly in the data collected by Linton et al. A complete discussion of the effects of all time lags across edges in Figure 1 is outside the scope of this paper.

## 2.3 Interventions

Data collected after a recent government intervention targeted to lower transmission of COVID-19 could produce a spuriously increased CFR. One primary tool of governments is the imposition of social-distancing measures, which decrease the amount and initial dose of exposures, thereby lowering  $S \rightarrow E$ . One incubation period after such a measure is enacted, the number of new cases will decrease, but the number of new deaths will not, since these deaths will be from cases diagnosed before the government intervention. This will upwardly bias the CFR for a few weeks after the intervention.

As in other pandemic influenzas, there may be a direct biological effect of increasing the infectious dose, leading to higher fatality rates (Paulo et al., 2010). Given an effective government intervention like social distancing, the infectious dose would decrease, directly lowering  $U \rightarrow UF$  and  $D \rightarrow DF$  and increasing  $U \rightarrow UR$  and  $D \rightarrow DR$ . This will upwardly bias current CFR estimates for some weeks after the initial intervention, since new deaths will still occur from cases whose onset time was before the intervention. The Centre for Evidence Based Medicine has a helpful page dedicated to COVID-19 viral dynamics like these (Heneghan et al., 2020).

Interventions to improve the quality of medical care can cause a drastic decrease in CFR, particularly when treatment options (e.g., drugs, blood transfusions, ventilators) become available or if training of healthcare workers (HCWs) improves. The effect can be highly pronounced in developing countries (Siddique, 1994) and is the subject of active study today (Hsiang et al., 2020; Warne et al., 2020). By the same logic as above, these interventions can lead to a spuriously increased CFR estimate. However, interventions that improve accessibility of medical care, such as the new health facilities being constructed around the world (Lardieri, 2020; J. Wang et al., 2020), can also increase testing and reporting. This increase will likely result in better data in the long term due to higher ascertainment of mild cases, although we have no data to support this conjecture.

## 2.4 Group characteristics

It is already well understood that certain groups have a higher CFR than other groups. In other words, the edges  $D \rightarrow DF$ ,  $D \rightarrow DR$ ,  $U \rightarrow UF$ , and  $U \rightarrow UR$  will have different values based on the characteristics of the sampled population, which could cause bias in either direction when estimating CFR. For example, the risk of death may be 34 to 73 times lower in people under 65 years old compared to those over 65 (Ioannidis et al., 2020). Furthermore, the incidences of comorbidities such as obesity, heart disease, smoking, genetics, and diabetes correlate with nation, socioeconomic status, race, sex, and more (Cai, 2020; Lee et al., 2014; Sliwa et al., 2008). In the context of surveillance data, without knowing the proportion of these groups in the sampling frame, which may not be uniform in time, the CFR can be biased in either direction. Chin et al. argue that incorporating county-level data about these covariates can result in a more equitable public-health response (Chin et al., 2020).

## 2.5 Imperfect reporting and attribution

Both the definition of a “case” and also the criteria under which an individual is eligible for testing can bias CFR estimates. Case definition, even within one nation, can change case counts dramatically. On February 12th, for example, the Chinese government changed the definition of “confirmed case” to include symptom-based diagnoses, resulting in a 600% increase in cases that day (Worldometer, 2020). Without information on how deaths were attributed beforehand, we do not know the magnitude of this bias. Serious problems have been introduced by poor reporting on behalf of governments. For example, the Johns Hopkins GitHub stopped providing surveillance data on recovered cases within the United States, because they argued the quality of the data was too low (CSSEGISandData, 2020). Furthermore, because testing is often reserved for the most severe cases (see the referenced Remarks from the White House),  $S2 \rightarrow D$  is inflated while  $AC \rightarrow D$  is deflated (United States Federal Government, 2020). This will spuriously increase  $E_{\text{naive}}$ . Evidently, detailed knowledge of how cases, deaths, and recoveries are defined and reported are prerequisite to understanding these biases, even if it will be impossible to correct for them without finer-grained



data.

Sensitivity and specificity of COVID-19 tests certainly affect all of  $AC, S1, S2, S3 \rightarrow D$  and  $AC, S1, S2, S3 \rightarrow U$ . A diagnostic test with a high false discovery rate will increase  $S \rightarrow S1$ , incorrectly inflating the denominator of  $E_{\text{naive}}$  and spuriously decreasing the CFR. Nonetheless, assays have improved with time; the initial test developed by the US Centers for Disease Control was ineffective (Sharfstein et al., 2020). Still, the serology assay used by Bendavid et al. (2020) had a putative sensitivity of 80% and specificity of 99.5%, which may still be too low to provide estimates of a small prevalence.

Distinctly from under-ascertainment, reporting of infectious disease by healthcare providers in the United States is often incomplete and normally has a mean time delay of 12-40 days depending on the pathogen (Jajosky & Groseclose, 2004). This means edges  $D \rightarrow RC$ ,  $DF \rightarrow RF$ , and  $DR \rightarrow RR$  are not 1.0. Because deaths are more likely to be reported by healthcare providers than confirmed cases or recoveries, ignoring time delay,  $D \rightarrow RC$  is less than  $DF \rightarrow RF$ , biasing  $E_{\text{naive}}$  upward. Depending on the relative time delays across these edges, estimators may be biased in either direction. For example, if  $D \rightarrow RC$  is lagged more than  $D \rightarrow RD$ , it would bias  $E_{\text{naive}}$  downward during the growth phase of an epidemic. To our knowledge, these time delays, which occur on a hospital-by-hospital basis, have not been quantified, and it is not obvious in what direction they will skew. The magnitude of this bias could be quite large for COVID-19. On Friday, April 17th, the Wuhan government reported 1,290 new fatalities, increasing their cumulative death toll by 50% in one day. They claimed the revision was because “medical workers . . . might have been preoccupied with saving lives, and there existed delayed reporting, underreporting, or misreporting.” This is a salient example of a high-magnitude bias from reporting errors that we can not correct, since we do not know at what time those deaths truly occurred (Kuo, 2020).

Evidence from past epidemics also indicates this bias may be significant for COVID-19. Even for severe illnesses such as Hepatitis C, healthcare providers can underreport cases by up to 12x (Klevens et al., 2014). Historically, the magnitude of underreporting depends heavily on ease of reporting for HCWs (e.g., electronic vs paper systems) and also mandatory reporting laws (Chorba et al., 1989; Panackal et al., 2002).

Finally, although survivorship bias may be small, misattribution of deaths (i.e., increased weight of  $S3 \rightarrow D$ ) may be significant. A recent JAMA article argued that COVID-19 positive patients with cardiac injury have a relative risk of death of 4.26 compared to those with no cardiac injury. Most of those patients also had abnormal electrocardiograms (Shi et al., 2020). Another case study described a healthy 53-year-old woman who tested positive for COVID-19, did not show any respiratory involvement, but developed acute myopericarditis with systolic dysfunction (Inciardi et al., 2020). Kidney involvement has also been found (Ronco & Reis, 2020). It is unclear how deaths in the presence of multiple diagnoses are being counted, and indeed, to which disease they should be attributed. Disentangling these relationships may be possible with regression on high-resolution clinical data. However, we have not seen this level of detail reflected in surveillance data. Comparisons with historical mortality data suggest tens of thousands of deaths are misattributed or unreported (J. Wu et al., 2020).

### 3 Naïve estimators

We access publicly available data courtesy of Johns Hopkins University, consisting of time-series data of recoveries, deaths, and confirmed cases stratified across several dozen groups (in this case,

primarily geographic locations) (Dong et al., 2020). Our computations were performed on April 18, 2020. We denote cohorts or groups of cases by indices  $g$ , belonging to a set  $G$ . For example,  $g$  could be “people under 60 years of age,” or “people in Wuhan.” For time points  $t = 1, 2, \dots, T = 41$ , we collect daily data as follows: for each group  $g \in G$  we collect  $R_t^g$ ,  $D_t^g$ , and  $C_t^g$ , which correspond to the number of new recoveries, new deaths, and new cases reported on day  $t$  within group  $g$ . We drop the group superscript  $g$  for population quantities:

$$R_t := \sum_{g \in G} R_t^g, \quad D_t := \sum_{g \in G} D_t^g, \quad C_t := \sum_{g \in G} C_t^g. \quad (1)$$

### 3.1 An estimator based on dividing deaths by cases

In early March 2020, the WHO estimate of the CFR, 3.4% was widely reported (Stelter, 2020; Ghebreyesus, 2020). This estimate is obtained from a naïve estimator;<sup>2</sup> specifically, the raw proportion of deaths among confirmed cases. Formally, as of March 6, 2020,

$$E_{\text{naive}} = \frac{\sum_t D_t}{\sum_t C_t} \approx 3.4\%. \quad (2)$$

As of April 18, 2020,  $E_{\text{naive}}$  was 6.9%. However, as we establish in Appendix A, in a setting without time delays, the naïve CFR is asymptotically unbiased for the true CFR if and only if the probability of reporting is the same for fatal and non-fatal cases. Moreover, it is unbiased in finite samples if and only if reporting is perfect. As discussed in Section 2, this is not true in the case of COVID-19. We also derive the finite-sample expectation of the estimator. Even asymptotically, the expectation of this estimator can become unboundedly far away from the true CFR as reporting goes to zero or the CFR goes to zero.

The naïve estimator requires no complex modeling or tuning parameters and is easy to interpret. As we argued in Section 2, there is no uniformly best method of measuring the CFR, and the naïve estimator should be viewed as one in a constellation of estimators giving a heuristic idea of the causal CFR. Nonetheless, the naïve estimator can be improved at little cost, and indeed, in this work, we suggest applying a simple correction for time-dependent reporting rates and alleviate two problems with the naïve estimator: time-lag between death and recovery, and time-dependence in the reporting rates of fatal and non-fatal cases.

### 3.2 An estimator based on observed outcomes

One can view the time lag in the numerator above (across the  $D \rightarrow DF$  link) as a consequence of “censoring” the data: a case has been identified, but the outcome is hidden. Methods for handling censored data have been studied for several decades in the statistical literature; in particular, in the context of the bootstrap (Efron, 1981). Although it is not the focus of our work, several others have already applied the bootstrap to COVID-19 data to find confidence intervals for other epidemiological parameters such as  $R_0$  (Read et al., 2020; Linton et al., 2020). This should also be done for the CFR for COVID-19, as Jewell et al. (2007) did for SARS, although the structure of the data used in that work differs from the current setting.

---

<sup>2</sup>To be clear, the exact form of their estimates is not made explicit in the WHO report.



	definition
$\psi_{t,g}$	probability of diagnosis, given death from COVID-19, onset at time $t$ , and membership in group $g$
$\phi_{t,g}$	probability of diagnosis, given recovery from COVID-19, onset at time $t$ , and membership in group $g$
$p_{t,g}$	probability of death, given onset at time $t$ , within group $g$
$\eta_t$	probability of death $t$ days after onset, given death occurs

**Table 1.** Parameters and variables relevant to our likelihood models

There is also a very simple estimator that avoids censoring by using only observed data, namely:

$$E_{\text{obs}} = \frac{\sum_{t=1}^T D_t}{\sum_{t=1}^T D_t + \sum_{t=1}^T R_t} \approx 20.7\%. \quad (3)$$

The CFR calculated by this estimator is upwardly biased, and we will briefly discuss why. This estimator accounts for the inflation of the denominator in the naïve estimator via the relative time lag between  $D \rightarrow RC$  and  $D \rightarrow DF$ . However, it assumes we observe the same fraction of recovered cases and fatal cases at the time of estimation. Thus, it has introduced a new bias, the relative reporting rate and time lag between  $D \rightarrow DF$  and  $D \rightarrow DR$ . We formalize the asymptotic inferential target of this estimator in Appendix B. Note that in all cases,  $E_{\text{obs}} \geq E_{\text{naive}}$ . In fact,  $E_{\text{obs}}$  is exactly  $3E_{\text{naive}}$  on April 18th. This large discrepancy is due to under-reporting of recoveries, specifically within the United States, as we note in Section 2. The United States has, as of April 19, roughly 40,000 deaths and 70,000 recoveries (Dong et al., 2020). Meanwhile, Spain has 20,000 deaths and 80,000 recoveries. Clearly, the reporting of recoveries in both nations is infrequent, and in the United States, it may be more than doubly so. The estimator  $E_{\text{obs}}$  illustrates the dangers of correcting one of many biases without considering total error. The estimator  $E_{\text{naive}}$  and the estimator we present in the next section do not use this recovery data.

## 4 Likelihood models

In this section, we describe a parametric model that, with respect to several strong modeling assumptions, accounts for two biases: time-varying reporting and disease-delayed mortality. For definitions and discussion of our model parameters, see Table 1. With reference to Figure 1, the model accounts for the time-dependence of  $D \rightarrow DF$  and from  $D \rightarrow DR$  (i.e., it models how the values of these conditional probabilities change as a function of time), and also for the time delay across those same edges. This model was previously used by Reich et al. (2012) for CFR estimation of influenza. It is a covariate-independent reporting model that assumes all non-fatal cases eventually recover, so it does not utilize the time series of recoveries. Similar parametric models have been used for CFR estimation during other pandemics (Ejima et al., 2012; Frome & Checkoway, 1985). When none of the biases in Section 2 other than the time dependence and time delay across  $D \rightarrow DF$  and  $D \rightarrow DR$  are large and the mathematical assumptions stated below are satisfied, this estimator has a smaller total error than  $E_{\text{naive}}$  and  $E_{\text{obs}}$ , evidenced by empirical evaluations in Reich et al. (2012).

Suppose that an individual is in group  $g$  and has infection onset at time  $t_{\text{on}}$ . Such a case has three possible outcomes, whose probabilities we define below (4). For further information, see also Table (2). First, the individual may eventually recover and be diagnosed. This occurs with probability  $\rho_{t_{\text{on}},g}^{(1)}$ , see Eq. (4a). Secondly, they may eventually die, having been diagnosed. This

	diagnosed	undiagnosed
death	included in scenario 1	included in scenario 3
recovery	included in scenario 2	included in scenario 3

**Table 2.** Outcome scenarios for COVID-19 patients, for patients with onset at time  $t_{\text{on}}$  and in group  $g$ . In our notation, scenario  $i$  occurs with probability  $\rho_{t_{\text{on}},g}^{(i)}$  and has count  $N_{t_{\text{on}},g}^{(i)}$ , for  $i = 1, 2, 3$ .

occurs with probability  $\rho_{t_{\text{on}},g}^{(2)}$ ; see Eq. (4b). Finally, they may go entirely undiagnosed. This occurs with the remaining probability,  $\rho_{t_{\text{on}},g}^{(3)}$ ; see Eq. (4c). In summary:

$$\rho_{t_{\text{on}},g}^{(1)} = \phi_{t_{\text{on}},g} (1 - p_{t_{\text{on}},g}), \quad (4a)$$

$$\rho_{t_{\text{on}},g}^{(2)} = \psi_{t_{\text{on}},g} p_{t_{\text{on}},g}, \quad (4b)$$

$$\rho_{t_{\text{on}},g}^{(3)} = 1 - \rho_{t_{\text{on}},g}^{(1)} - \rho_{t_{\text{on}},g}^{(2)} = p_{t_{\text{on}},g} (1 - \psi_{t_{\text{on}},g}) + (1 - p_{t_{\text{on}},g})(1 - \phi_{t_{\text{on}},g}). \quad (4c)$$

Accordingly, at each onset time  $t_{\text{on}}$  and for each group  $g$ , there are  $N_{t_{\text{on}},g}^{(1)}$ ,  $N_{t_{\text{on}},g}^{(2)}$ , and  $N_{t_{\text{on}},g}^{(3)}$  individuals who eventually recover, die, or go undiagnosed respectively. Given a total number of cases within group  $g$  with onset at time  $t_{\text{on}}$ , denoted  $N_{t_{\text{on}},g}^*$ , we model the outcomes via a *multinomial model*:

$$(N_{t_{\text{on}},g}^{(1)}, N_{t_{\text{on}},g}^{(2)}, N_{t_{\text{on}},g}^{(3)}) \stackrel{\text{ind.}}{\sim} \text{Multinomial}(N_{t_{\text{on}},g}^*, \rho_{t_{\text{on}},g}^{(1)}, \rho_{t_{\text{on}},g}^{(2)}, \rho_{t_{\text{on}},g}^{(3)}), \quad \text{for all } t_{\text{on}} \text{ and } g. \quad (5)$$

We assume these are independent across onset times and group. Furthermore, we assume knowledge of certain horizon probabilities. In order to define an estimator, we also need probabilities  $\eta_{t,t_{\text{on}},g}$ , for  $t \geq 0$ . These are probabilities that, given an individual is in group  $g$  and has onset of infection at time  $t_{\text{on}}$ , they die  $t$  days later. We make the assumption that  $\eta_{t,t_{\text{on}},g} \equiv \eta_t$  for all  $t_{\text{on}}, g$ . That is, these probabilities are time- and group-invariant. See Reich et al. (2012) for further analysis and evaluation of this model.

**The estimator:** Let  $N_{t_{\text{on}},g}$  denote the *reported* total number of cases of COVID-19 with onset at time  $t_{\text{on}}$  in group  $g$ . Unfortunately, this is not the quantity of true interest. Instead, as mentioned above, we need  $N_{t_{\text{on}},g}^*$ , which is the number of number of both reported and *unreported* cases. Let  $\mathbf{E}$  denote the expectation operator. In particular,  $\mathbf{E}_{N_{t_{\text{on}},g}^*}$  will be an expectation with respect to the multinomial model in Equation 5 indexed by  $N_{t_{\text{on}},g}^*$ . As a simplifying assumption, we assume  $\rho_{t_{\text{on}},g}^{(2)} \equiv p_g$ ; that is, the group-specific death probability or CFR is time-invariant. If the  $p_g$  are small, then  $N_{t_{\text{on}},g}^* \approx N_{t_{\text{on}},g} / \phi_{t_{\text{on}},g}$ , in which case from our multinomial (5), it is easy to check that:

$$\mathbf{E}_{N_{t_{\text{on}},g}^*} [N_{t_{\text{on}},g}^{(2)}] = N_{t_{\text{on}},g}^{(2)} \rho_{t_{\text{on}},g}^{(2)} \approx N_{t_{\text{on}},g} \frac{\rho_{t_{\text{on}},g}^{(2)}}{\phi_{t_{\text{on}},g}}. \quad (6)$$

In particular, if we assume that death is a rare event, then a Poisson approximation will be accurate:

$$N_{t_{\text{on}},g}^{(2)} \sim \text{Poisson} \left( N_{t_{\text{on}},g} \frac{\rho_{t_{\text{on}},g}^{(2)}}{\phi_{t_{\text{on}},g}} \right), \quad (7)$$

where  $N_{t_{\text{on}},g}$  denotes the number of cases with onset at time  $t_{\text{on}}$  within group  $g$ . In view of (4b), this may be rewritten as:

$$N_{t_{\text{on}},g}^{(2)} \sim \text{Poisson} \left( N_{t_{\text{on}},g} \frac{\psi_{t_{\text{on}},g} p_{t_{\text{on}}}}{\phi_{t_{\text{on}},g}} \right). \quad (8)$$

If we make either an assumption that the reporting rates are group-invariant, or that there is perfect fatal-case reporting,  $\psi_{t,g} \propto \phi_{t,g}$ , then it is possible to rewrite the model in the form:

$$\log \left\{ \mathbf{E} \left[ N_{t_{\text{on}},g}^{(2)} \right] \right\} \approx \log \{ N_{t_{\text{on}},g} + \beta_0 + \alpha_{t_{\text{on}}} + \gamma_g \}, \quad (9)$$

where  $\beta_0$  is a proportionality constant,  $\gamma_g$  is a group-specific parameter (the relative CFR), and  $\alpha_{t_{\text{on}}}$  is a time-specific parameter. Finally, given these values, along with the death probabilities  $\eta_t$ , an expectation-maximization scheme can be carried out to compute a maximum-likelihood estimator. For further details, see sections 3.2 and 3.3 of Reich et al. (2012). They show empirically that as long as  $p_g$  stays below 0.05, and their assumptions are approximately satisfied, the estimated CFR has relative error  $< 0.1$  as compared to the ground truth. Their results also indicate that this model is insensitive to various misspecifications, including the distribution of deaths,  $\eta_t$ . We confirm this in Figure 2 by sampling parameters of  $\eta_t$  from their estimated confidence intervals (Linton et al., 2020).

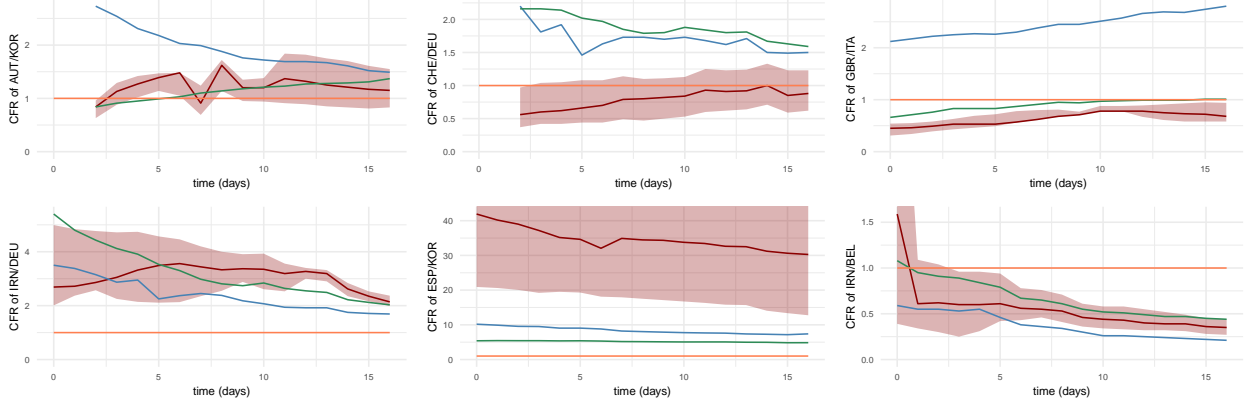
## 5 Results

We report the results of our analysis of open-sourced COVID-19 data from Johns Hopkins, under the assumption that the reporting rates  $\psi_t$  and  $\phi_t$  are group-invariant. We contribute an open-source multithreaded implementation of Reich et al. and a plotting utility that will allow reproducibility of these results, as shown in Figure 2. Finally, we report the relative CFR of women to men in Germany and Belgium using sex-disaggregated data from Riffe (2020).

### 5.1 Estimates of Relative CFRs

The corrected relative CFRs, calculated for six combinations of nations, are listed in Figure 2. In some cases, such as the comparison between England (GBR) and Italy (ITA), our estimator flips the direction of the relative CFR. In other words,  $E_{\text{naive}}$  and  $E_{\text{obs}}$  suggest that England has a higher CFR than Italy, while  $E_{\text{Reich}}$  suggests otherwise. The same effect happens in the case of Switzerland (CHE) vs Germany (DEU), with an additional shrinkage in the distance towards 1, indicating the relative CFR is more similar than  $E_{\text{naive}}$  and  $E_{\text{obs}}$  would suggest. The estimate of the relative CFR for Spain (ESP) to South Korea (KOR) predicted by  $E_{\text{Reich}}$  is high at 30.27. Although we assumed in Section 4 that the relative CFR is constant in time, we report our results as a time series in Figure 2. We obtain this time series by calculating results as if we had run our estimator on every day between April 2, 2020 and April 16, 2020 using the cumulative data.

Using the data from Riffe (2020), we calculated the relative CFR of women to men in Germany and Belgium. In Germany,  $E_{\text{naive}} = 1.51$  and  $E_{\text{Reich}} = 1.14$  (Sensitivity 1.14 – 1.22). In Belgium,  $E_{\text{naive}} = 1.68$  and  $E_{\text{Reich}} = 1.25$  (Sensitivity 1.13 – 1.26). Time series data of recoveries is not available, so we could not calculate  $E_{\text{obs}}$ . We chose Germany and Belgium because the data from these nations had about two months of seemingly reliable, day-by-day, sex-disambiguated data that roughly matched the numbers from Johns Hopkins. The dataset from Riffe et al. was still under development at the time we ran these estimates.



Estimator	AUT/KOR	CHE/DEU	GBR/ITA	IRN/DEU	ESP/KOR	IRN/BEL
$E_{naive}$	1.37	1.59	1.01	2.03	4.96	0.44
$E_{obs}$	1.49	1.5	2.8	1.69	7.28	0.21
$E_{Reich}$	1.15	0.81	0.68	2.03	30.27	0.38
	(0.83, 1.55)	(0.62, 1.23)	(0.58, 0.94)	(1.99, 2.37)	(12.75, 55.23)	(0.27, 0.44)

**Figure 2.** The estimators  $E_{naive}$ ,  $E_{obs}$ , and  $E_{Reich}$  presented as time series from April 2, 2020 to April 16, 2020. Our estimator, in red, implements the correction for time-dependent relative reporting rates between countries identified by their ISO abbreviations. Sensitivity of our results to misspecification of  $\eta_t$  is reported by setting  $\eta_t$  to be a discretized Gamma distribution with mean (12.8, 17.5) and variance (5.2, 9.1), the lower and upper extremes of the 95% confidence intervals referenced (Linton et al., 2020). The ribbon shows the maximum and minimum values of the estimator  $E_{Reich}$  at each time point under any combination of these conditions. The expectation maximization algorithm converged in all cases with negligible variance. We include the relative CFR of Spain to South Korea as an example of two countries for which our assumptions are particularly badly violated. Consequently, the method is unstable in that case (although we have no ground truth data for confirmation). Notice each plot has a different vertical axis scaling. We have included an orange line at a relative CFR of 1 to indicate the point when two countries have the same estimated CFR; this provides a reference point between the plots.

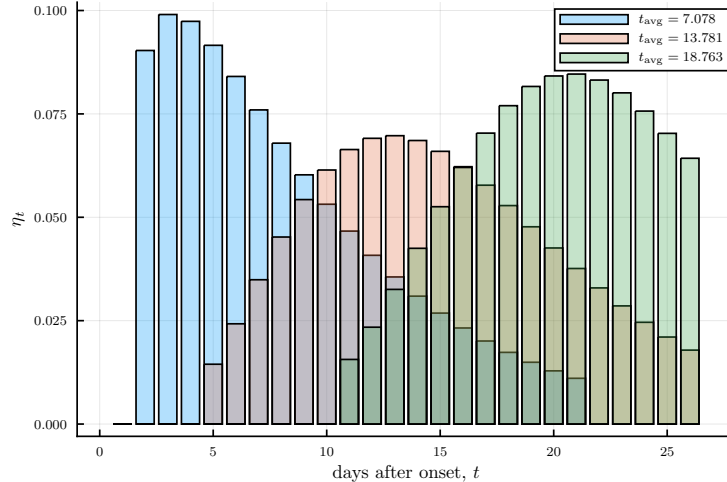
## 5.2 Choosing $\eta_t$

As described in Section 4, we assume access to probabilities  $\eta_t$  that indicate the probability of death occurring for a fatal case  $t$  days post-onset of COVID-19. Since our model indexes time by day, we need to set  $\eta_t$  for integers  $t \geq 0$ . Our choice of distribution is the best-fit discretized Gamma distribution to the fatality time horizons from Chinese data (shape parameter  $k = 4.726$  and scale  $\theta = 3.174$ ) (Linton et al., 2020). These parameters were roughly consistent across several other studies (Mizumoto & Chowell, 2020; W. Wang et al., 2020). We discretized the probability density function  $\eta_t$  to the days  $t = 0, \dots, T$ . Formally, after selecting a mean parameter  $t_{avg} > 0$ , we determine the probabilities  $\eta_t$  by<sup>3</sup>

$$\eta_t \propto t^{k-1} e^{-t/\theta}, \quad t = 0, \dots, T. \quad (10)$$

Stated differently, for a given mean parameter  $t_{avg}$ , we define a probability measure  $\eta \in \mathbf{R}_+^T$ , on  $t \in \{0, \dots, T\}$ , according to (10). See Figure 3 for an illustration of this distribution. In our

<sup>3</sup>The notation  $\propto$  is hiding a positive normalization constant to make  $\eta$  a probability measure.



**Figure 3. Illustration of distribution of death times for fatal cases,  $\eta_t$ , in days post onset.** Our method assumes knowledge of the probability of death for a fatal case  $t$  days post-onset of COVID-19. This data was estimated by fitting a Gamma distribution to the fatality time horizons from Chinese data in early February, 2020 (Linton et al., 2020). We discretize the distribution by day and also truncate it to 25 days long, both for numerical stability and also because very few deaths occurred past this point in the real data. The mean time to death was 15.0 (95% CI 12.8 – 17.5). The standard deviation was 6.9 (95% CI 5.2 – 9.1), which we also used in our sensitivity analysis above. The three separate gamma distributions plotted above have different choices of mean value for illustrative purposes, to show the qualitative effect of changing the parameter.

experiments, we truncate at  $T = 25$ , both for numerical stability and also because very few deaths occur after 25 days in the data used to fit the Gamma distribution.

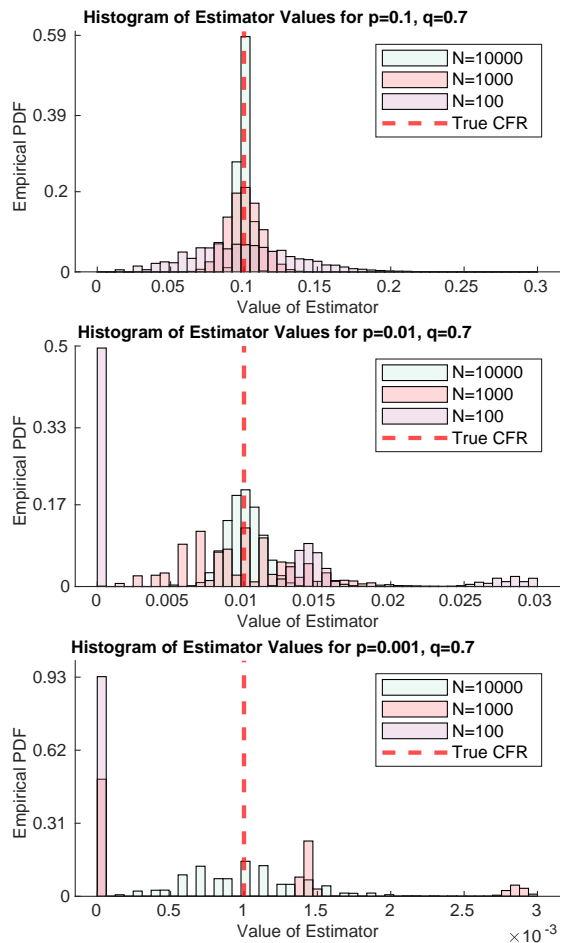
## 6 Discussion

We emphasize again that the procedure that we have presented for estimation of relative CFR seeks to address only a subset of the biases that impinge upon the ascertainment of this important population-level parameter. We explicitly account for the time-dependence of reporting rates that may differ between fatal and non-fatal cases. We have separate time-dependent reporting rates for cases that will eventually be fatal or non-fatal, addressing the fact that reporting is higher among severe cases. Deaths are known to vary with some combination of healthcare quality and age, which can be quantified with a relative CFR estimate. Our covariate-independent reporting rate assumption likely does not hold in practice. Indeed, the relative CFR of Spain with respect to Korea (two countries whose time-dependent reporting rates are probably different) yields a value of 30.27, likely speaking to the limitations of this method, although we do not have ground truth. Although Reich et al. present extensive experimental evaluations and some theory indicating that the method outperforms  $E_{\text{naive}}$  under given modeling assumptions, it is not generally possible to check how closely these assumptions hold, due to overparameterization of the unrestricted model. This issue may be mitigated by working with domain experts who understand each group’s sampling and reporting patterns. Another issue is that our estimator uses parameters  $\eta_t$  that are not estimated strictly from surveillance data but rather from individualized death times (Linton et al., 2020).

We believe that the maximum-likelihood estimator that we have presented may provide a more valid correction of relative reporting rates between German women and men rather than between South Korean people and Italian people, given that reporting rates by sex may be closer to identical than reporting rates by country, although biases by sex still exist (Guerra-Silveira & Abad-Franch, 2013). Demographers have argued that releasing data stratified by sex, age, and other demographic groups would aid in understanding the spread and fatality rates of COVID-19 (Dowd et al., 2020). Although certain teams like Riffe (2020) are currently assembling this data, many agencies are reporting such strata infrequently or not at all, making data collection difficult. To our knowledge, there is no well-established data repository (like the Johns Hopkins repository) that contains time-series data of deaths and cases stratified by sex, age, etc.

Many of the key biases that we reviewed in Section 2 remain unaddressed in current data-collection and data-analysis pipelines. Variations in the nature of the population within the sampling frame that gets tested, due to government- or geography-specific protocols will cause any CFR estimate to be unreliable. In particular, details in the definitions of terms across countries and times can result in severe bias in time series data; for example, China’s explicit policy was that they would not report asymptomatic cases until April 1, 2020, when the policy changed (Jiang, 2020). Accounting for many of the biases we have discussed may be possible with great effort by many data analysts. However, it is equally important for the statistical community to channel much of its energy into a unison clarion call to governments: to get estimates to support consequential policy-making, we need more and better data.

Contact tracing is a particularly powerful way to obtain data that allow otherwise intractable biases to be controlled, since it ex-



	$\delta/p = 0.1$	$\delta/p = 0.01$	$\delta/p = 0.001$
$q = 0.9$	$N = 1$	$N = 2$	$N = 3$
$q = 0.5$	$N = 4$	$N = 7$	$N = 10$
$q = 0.1$	$N = 22$	$N = 44$	$N = 66$

**Figure 4.** Under the assumptions in the contact-tracing scenario,  $E_{\text{naive}}$  converges in few samples to an unbiased estimator of the CFR. The bound on the number of samples  $N$  derived in the Appendix A,  $N \geq \lceil \frac{\log(\delta/p)}{\log(1-q)} \rceil$  (Eq. 16) was used to calculate the values in the table. Notice  $N$  is a function of  $\delta/p$  (the acceptable relative error) and  $q$  (the reporting rate). The empirical distribution functions of  $E_{\text{naive}}$  with different parameters of  $p$  and  $N$  and a reporting rate  $q = 0.7$  are plotted. Notice that as  $p$  decreases, detecting a case will require larger  $N$ .



pands the sampling frame to include a much larger portion of our target population, specifically mild cases. Contact tracing is the process of reaching out to all individuals (“contacts”) who were recently exposed to a known COVID-19 positive individual, removing them from circulation, and monitoring their health. The same is done for contacts of contacts, and so on, for an appropriate number of iterations. We suggest that all of these contacts should be tested for COVID-19 one incubation period after exposure, regardless of whether or not they are symptomatic. The number of data points gleaned from this strategy will be lower than the number of data points from surveillance data. However, the population sampled using this strategy would be closer to the target population, since it would include asymptomatic cases. Furthermore, there is no issue with time lag, since these cases can be tracked systematically. Specifically, assume the non-response rate to contact tracing is identical for asymptomatic and symptomatic cases. As we prove in Appendix A, this is the exact condition under which  $E_{\text{naive}}$  is an asymptotically unbiased estimator. Moreover, the estimator has desirable finite-sample properties in such a setting. Letting  $p$  be the true CFR and  $q$  be the reporting rate among infected cases, we have that  $E_{\text{naive}}$  lies within a range  $\delta$  of  $p$  in  $N = \lceil \frac{\log(\delta/p)}{\log(1-q)} \rceil$  samples; see Eq. 16 below. As seen in Figure 4,  $N$  does not need to be too large to insure that the bias of  $E_{\text{naive}}$  is small, although with small  $p$ , sampling any fatal cases requires  $N$  to be on the order of  $\frac{1}{p}$  in this simplified model.

Contact tracing does not eliminate all biases. The assumption that non-response rates do not vary by case severity will not hold unless responses are mandatory, possibly introducing significant error, especially as  $p$  becomes small. Assay sensitivity and specificity may still cause errors. Most importantly, care must be taken to make valid inferences about the desired target population based on individualized contact tracing data that may come from a restricted sample. One major hurdle is estimation of  $p$  when it is small: as shown in Figure 4, if death is an very rare event,  $N$  would need to be large in order to ensure enough fatal cases are sampled. Finally, such data may be easier to collect and release in some countries and jurisdictions than others. For example, within the United States, medical privacy and consent laws may make it difficult to ever test a truly random sample of the population, or to release the fine-grained data necessary for corrected estimators. These challenges, outside the scope of our work, are well studied in the field of survey sampling and reweighting.

## Acknowledgments

A. N. A. was partially supported by the National Science Foundation Graduate Research Fellowship Program. R. P. was partially supported by a UC Berkeley University Fellowship via the ARCS Foundation. We wish to thank Constance Angelopoulos for illustrating Figure 1, Esther Rolf and Ilija Radosavovic for reading and commenting on the manuscript, [Hypernet](#) for providing compute resources, and Anthony Ebert for contributing code and comments to our GitHub pre-release ([see his Git here](#)). Finally, we thank the editor and reviewers of the Harvard Data Science Review for providing valuable feedback.

## A Derivation of the expectation of $E_{\text{naive}}$

In this section we derive the expectation of  $E_{\text{naive}}$ . Our derivation will employ a stripped-down notation, since here we deal with individual random variables for each COVID-19 infected person instead of time-series data. Index the infected population with the integers  $\{1, \dots, N\}$ . Let  $T_i \sim$

$\text{Ber}(p)$  be a Bernoulli random variable representing whether or not person  $i \in \{1, \dots, N\}$  died, and let  $W_i \sim \text{Ber}(q)$  be a Bernoulli random variable representing whether or not person  $i$  was diagnosed with the virus. We want to estimate  $p$ , but we only have the reported number of deceased patients with COVID-19,  $V_i = T_i W_i$ ,  $i \in \{1, \dots, N\}$ . Defining  $r := \text{Cov}(T_i, W_i)$ , the joint distribution of  $(T_i, W_i)$  can be expressed as a contingency table:

	$T_i = 1$	$T_i = 0$
$W_i = 1$	$r + pq$	$q(1 - p) - r$
$W_i = 0$	$p(1 - q) - r$	$1 + r - p - q + pq$

Several of the results discussed in the main paper follow as simple consequences of the calculation of the expectation of  $E_{\text{naive}}$ : (1)  $E_{\text{naive}}$  is unbiased for  $p$  in finite samples if and only if  $q = 1$ ; (2)  $E_{\text{naive}}$  is unbiased for  $p$  as  $N \rightarrow \infty$  if and only if  $r = 0$ ; (3) if there is an  $\epsilon > 0$  error in the estimation of  $r$ , for example due to incorrect attribution of fatalities to COVID-19, then  $E_{\text{naive}}$  has unbounded expectation  $q \rightarrow 0$  and unbounded relative error as  $p \rightarrow 0$ ; and (4) if  $r = 0$ , the smallest  $N$  such that  $|\mathbf{E}[E_{\text{naive}}] - p| \leq \delta$  is  $N = \lceil \frac{\log(\delta/p)}{\log(1-q)} \rceil$ .

The distribution of  $V_i$  is Bernoulli with  $\mathbf{P}[V_i = 1] = r + pq$ . Also define  $\gamma_1 = \mathbf{P}[W_i = 1 | T_i = 1] = (r + pq)/p$ . Applying the tower property of conditional expectation and using the exchangeability of the  $(T_i, W_i)$  pairs, we have:

$$\mathbf{E} \left[ \frac{\sum_{i=1}^N V_i}{\sum_{j=1}^N W_j} \right] = \sum_{i=1}^N \mathbf{E} \left[ \frac{V_i}{\sum_{j=1}^N W_j} \right] = N \mathbf{E} \left[ T_1 \mathbf{E} \left[ W_1 \frac{1}{W_1 + \sum_{l=2}^N W_l} \middle| T_1 \right] \right]. \quad (11)$$

Since  $W_1$  is independent of  $W_2, \dots, W_N$ , we can express the sum in the denominator as a binomial random variable,  $B \sim \text{Bin}(N - 1, q)$ , since it is a sum of  $N - 1$  i.i.d. Bernoulli random variables  $(W_2, \dots, W_N)$  with parameter  $q$ . Note the fact that  $\mathbf{E} \left[ \frac{1}{1+B} \right] = ((1 - (1 - q)^N))/Nq$ . Then, evaluating the innermost expectation first:

$$N \mathbf{E} \left[ T_1 \mathbf{E} \left[ W_1 \frac{1}{W_1 + B} \middle| T_1 \right] \right] = N \mathbf{E} \left[ T_1 \gamma_1 \mathbf{E} \left[ \frac{1}{1+B} \right] \middle| T_1 \right] = p \frac{\gamma_1}{q} (1 - (1 - q)^N). \quad (12)$$

Finally, substituting for  $\gamma_1$ , we obtain the final form:

$$\mathbf{E}[E_{\text{naive}}] = \frac{r + pq}{q} (1 - (1 - q)^N). \quad (13)$$

Recall that  $q$  is the probability of reporting given an infection, and  $p$  is the probability of death given an infection. Since  $1 - q = 0$  implies  $r = 0$  because  $W$  becomes deterministic,  $E_{\text{naive}}$  is unbiased for  $p$  if and only if  $q = 1$ . Moreover, if  $q \neq 1$ , taking  $N \rightarrow \infty$  shows  $E_{\text{naive}}$  is asymptotically biased for  $p$  if and only if  $r = 0$ . Both of these conditions are violated for any real disease. Interestingly this empirical CFR is not constrained to be an underestimate, and can overestimate  $p$  if  $p \leq \frac{r(1 - (1 - q)^N)}{q(1 - q)^N}$ .

Under the assumption  $r \geq \epsilon$ , the (asymptotic) overestimate can be unboundedly bad. This may arise if there is an  $\epsilon$  error in estimating the covariance (which we assume to be nonnegative), because some people are diagnosed with COVID-19 but their death is not *caused* by COVID-19. In this context:

$$\lim_{q \rightarrow 0} \frac{r + pq}{q} \geq \lim_{q \rightarrow 0} \frac{\epsilon + pq}{q} = \infty. \quad (14)$$

In other words, as the rate of reporting,  $q$ , decreases or the covariance between death and reporting increases, the CFR estimate gets worse, ultimately becoming infinitely bad, as long as there is a spuriously positive relationship between death and diagnosis. Similarly, the ratio  $E_{\text{naive}}/p$  can become infinitely bad as the product  $pq$  decreases. Note that if there is no spurious relationship,  $r \rightarrow 0$  as  $p \rightarrow 0$  or  $q \rightarrow 0$  since  $W$  and  $T$  become deterministic under those conditions. In the case of COVID-19, neither  $q$  nor  $p$  are near zero, but the limiting case helps to exhibit the qualitative performance of the estimator—it gets more bias-prone with smaller  $p$  and  $q$  and with larger  $r$ .

Finally, assume that  $W$  and  $T$  are independent so  $r = 0$ . Then,  $|\mathbf{E}[E_{\text{naive}}] - p| \leq \delta$  implies that

$$p| - (1 - q)^N | \leq \delta \implies N \log(1 - q) \leq \log(\delta/p) \implies N \geq \frac{\log(\delta/p)}{\log(1 - q)}. \quad (15)$$

Constraining  $N$  to be the smallest  $N \in \mathbf{N}$  such that  $|\mathbf{E}[E_{\text{naive}}] - p| \leq \delta$  gives

$$N = \lceil \frac{\log(\delta/p)}{\log(1 - q)} \rceil. \quad (16)$$

## B Derivation of the asymptotics of $E_{\text{obs}}$

We borrow notation and proof technique from Reich et al. (2012). We use the same notation as the main paper. This proof applies to the group-independent reporting rate model. In Reich et al., a similar proof is shown that applies in the case of the constant-proportion assumption.

In addition to the notation in the main paper, define:  $d_{t,g} := \mathbf{E}[D_{t,g}] = N_{t,g}^* p g \psi_t$  and  $r_{t,g} := \mathbf{E}[R_{t,g}] = N_{t,g}^* (1 - p_g) \phi_t$ . Also, introduce two nonrandom functions,  $F_d(t) \rightarrow [0, 1]$  and  $F_r(t) \rightarrow [0, 1]$ , where  $F_d$  represents the fraction of confirmed, fatal cases who have died by time  $t$ . Similarly  $F_r$  represents the fraction of confirmed, non-fatal cases who have recovered by time  $t$ . During an active outbreak, we have  $F_d < 1$  and  $F_r < 1$ . Finally, define  $T$  as the current time; all sums over time below have an upper limit of  $T$  unless otherwise specified. We seek an asymptotic limit for:

$$E_{\text{obs}} = \frac{F_d(T) \sum_{t_2} D_{t_2,g}}{(F_d(T) \sum_{t_1} D_{t_1,g}) + (F_r(T) \sum_{t_1} R_{t_1,g})}. \quad (17)$$

By the weak law of large numbers,  $D_{t,g}$  and  $R_{t,g}$  converge to their expectations, so we have:

$$\frac{F_d(T) D_{t,g}}{N_{t,g}^*} \xrightarrow{p} \frac{F_d(T) d_{t,g}}{N_{t,g}^*} = F_d(T) p g \psi_t, \quad (18)$$

and similarly,

$$\frac{F_r(T) R_{t,g}}{N_{t,g}^*} \xrightarrow{p} \frac{F_r(T) r_{t,g}}{N_{t,g}^*} = F_r(T) (1 - p_g) \phi_t. \quad (19)$$

Now we focus on the denominator. We have to introduce a “smoothness” assumption: the number of infected people at each timestep,  $N_{t_1,g}^*$ , has a constant ratio with respect to the number of infected people at each other timestep,  $N_{t_2,g}^*$ . In particular,  $\lambda_{t_1,t_2,g}$  corresponds roughly to the growth rate of the disease. Although this quantity would vary based on many factors in a real setting, we assume it to be a constant here. In particular, as  $N_{t_1,g}^* \rightarrow \infty$  and  $N_{t_2,g}^* \rightarrow \infty$ ,

$$\frac{N_{t_1,g}^*}{N_{t_2,g}^*} \rightarrow \lambda_{t_1,t_2,g}. \quad (20)$$

Therefore, we have by Slutsky's theorem that:

$$\frac{F_d(T)D_{t_1,g} + F_r(T)R_{t_1,g}}{N_{t_2,g}^*} \xrightarrow{p} \lambda_{t_1,t_2,g}(F_d(T)p_g\psi_{t_1} + F_r(T)(1-p_g)\phi_{t_1}). \quad (21)$$

Now, applying the weak law of large numbers and our assumption:

$$\frac{F_d(T)D_{t_1,g} + F_r(T)R_{t_1,g}}{N_{t_2,g}^*} = \frac{N_{t_1,g}^*}{N_{t_2,g}^*} \left( \frac{F_d(T)D_{t_1,g}}{N_{t_1,g}^*} + \frac{F_r(T)R_{t_1,g}}{N_{t_1,g}^*} \right) \xrightarrow{p} \lambda_{t_1,t_2,g}(F_d(T)p_g\psi_{t_1} + F_r(T)(1-p_g)\phi_{t_1}). \quad (22)$$

Then, by Slutsky's theorem on the sum which is the denominator of  $E_{\text{obs}}$  divided by  $N_{t_2,g}^*$ ,

$$\Sigma_{t_1} \left[ \frac{F_d(T)D_{t_1,g} + F_r(T)R_{t_1,g}}{N_{t_2,g}^*} \right] \xrightarrow{p} \Sigma_{t_1} \lambda_{t_1,t_2,g}(F_d(T)p_g\psi_{t_1} + F_r(T)(1-p_g)\phi_{t_1}). \quad (23)$$

Now, considering one term from the sum over  $t_2$  in  $E_{\text{obs}}$ , we have:

$$\frac{F_d(T)D_{t_2,g}}{\Sigma_{t_1} F_d(T)D_{t_1,g} + F_r(T)R_{t_1,g}} = \frac{F_d(T)D_{t_2,g}/N_{t_2,g}^*}{\Sigma_{t_1} (F_d(T)D_{t_1,g} + F_r(T)R_{t_1,g})/N_{t_2,g}^*} \quad (24)$$

$$\xrightarrow{p} \frac{F_d(T)p_g\psi_{t_2}}{\Sigma_{t_1} (\lambda_{t_1,t_2,g}(F_d(T)p_g\psi_{t_1} + F_r(T)(1-p_g)\phi_{t_1}))}. \quad (25)$$

Finally, rearranging terms and appealing once more to Slutsky's theorem:

$$E_{\text{obs}} \xrightarrow{p} \Sigma_{t_2} \frac{F_d(T)p_g\psi_{t_2}}{F_d(T)(\Sigma_{t_1} \lambda_{t_1,t_2,g} p_g \psi_{t_1}) + F_r(T)(\Sigma_{t_1} \lambda_{t_1,t_2,g} (1-p_g)\phi_{t_1})}. \quad (26)$$

This is clearly a biased estimator of  $p_g$ .

## References

- Atkins, K. E., Wenzel, N. S., Ndeffo-Mbah, M., Altice, F. L., Townsend, J. P., & Galvani, A. P. (2015). Under-reporting and case fatality estimates for emerging epidemics. *British Medical Journal*, *350*, h1115.
- Battegay, M., Kuehl, R., Tschudin-Sutter, S., Hirsch, H. H., Widmer, A. F., & Neher, R. A. (2020). 2019-novel Coronavirus (2019-nCoV): estimating the case fatality rate—a word of caution. *Swiss Medical Weekly*, *150*(0506).
- Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., ... Bhattacharya, J. (2020). COVID-19 antibody seroprevalence in Santa Clara county, California. *medRxiv*. Retrieved from <https://www.medrxiv.org/content/early/2020/04/17/2020.04.14.20062463> doi: 10.1101/2020.04.14.20062463
- Cai, H. (2020). Sex difference and smoking predisposition in patients with COVID-19. *The Lancet Respiratory Medicine*, *8*(4), e20.
- Chin, T., Kahn, R., Li, R., Chen, J. T., Krieger, N., Buckee, C. O., ... Kiang, M. V. (2020). US county-level characteristics to inform equitable COVID-19 response. *medRxiv*.

- Chorba, T. L., Berkelman, R. L., Safford, S. K., Gibbs, N. P., & Hull, H. F. (1989). Mandatory reporting of infectious diseases by clinicians. *Journal of the American Medical Association*, *262*(21), 3018–3026.
- Chowell, G., Hyman, J. M., Bettencourt, L. M., & Castillo-Chavez, C. (2009). *Mathematical and statistical estimation approaches in epidemiology*. Springer.
- CSSEGISandData. (2020). *Johns Hopkins Center for Systems Science and Engineering Github: Upcoming changes in time series tables*. Published Online.
- Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*. (To appear.)
- Donnelly, C. A., Ghani, A. C., Leung, G. M., Hedley, A. J., Fraser, C., Riley, S., . . . others (2003). Epidemiological determinants of spread of causal agent of Severe Acute Respiratory Syndrome in Hong Kong. *The Lancet*, *361*(9371), 1761–1766.
- Dowd, J., Andriano, L., Brazel, D., Rotondi, V., Block, P., Ding, X., . . . Mills, M. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences*. (To appear.)
- Eames, K. T., & Keeling, M. J. (2003). Contact tracing and disease control. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1533), 2565–2571.
- Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, *76*(374), 312–319.
- Ejima, K., Omori, R., Cowling, B. J., Aihara, K., & Nishiura, H. (2012). The time required to estimate the case fatality ratio of influenza using only the tip of an iceberg: Joint estimation of the virulence and the transmission potential. *Computational and Mathematical Methods in Medicine*, *2012*.
- Fauci, A. S., Lane, H. C., & Redfield, R. R. (2020). *COVID 19: Navigating the uncharted* (Vol. 382). Massachusetts Medical Society.
- Frome, E. L., & Checkoway, H. (1985). Use of Poisson regression models in estimating incidence rates and ratios. *American Journal of Epidemiology*, *121*(2), 309–323.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164.
- Ghebreyesus, T. A. (2020, March 3). *WHO Director-General’s opening remarks at the media briefing on COVID-19*.
- Goodman, J. D., & Rothfeld, M. (2020). *1 in 5 New Yorkers may have had Covid-19, antibody tests suggest*. The New York Times. (Online news article. Surveys were performed at grocery stores by NY state.)
- Guerra-Silveira, F., & Abad-Franch, F. (2013). Sex bias in infectious disease epidemiology: Patterns and processes. *PLoS One*, *8*(4), e62390.

- Heneghan, C., Brassey, J., & Jefferson, T. (2020). *SARS-CoV-2 viral load and the severity of COVID-19*. Published Online. (Online resource)
- Hsiang, S., Allen, D., Annan-Phan, S., et al. (2020). The effect of large-scale anti-contagion policies on the Coronavirus (COVID-19) pandemic. *medRxiv*.
- Inciardi, R., Lupi, L., Zaccone, G., Italia, L., Raffo, M., Tomasoni, D., . . . others (2020). Cardiac involvement in a patient with Coronavirus disease 2019 (COVID-19). *Journal of the American Medical Association Cardiology*. (To appear.)
- Ioannidis, J. P., Axfors, C., & Contopoulos-Ioannidis, D. (2020). Population-level COVID-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters. *medRxiv*.
- Jajosky, R. A., & Groseclose, S. L. (2004). Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC Public Health*, 4(1), 29.
- Jewell, N. P., Lei, X., Ghani, A. C., Donnelly, C. A., Leung, G. M., Ho, L., . . . Hedley, A. J. (2007). Non-parametric estimation of the case fatality ratio with competing risks data: an application to Severe Acute Respiratory Syndrome (SARS). *Statistics in Medicine*, 26(9), 1982–1998.
- Jiang, S. (2020). *China to begin reporting asymptomatic cases in its daily tally*. Cable News Network, Published Online. (Online news article.)
- Jordan, M. I. (2004). Graphical models. *Statistical Science*, 19, 140–155.
- Klevens, R. M., Liu, S., Roberts, H., Jiles, R. B., & Holmberg, S. D. (2014). Estimating acute viral hepatitis infections from nationally reported cases. *American Journal of Public Health*, 104(3), 482–487.
- Kuo, L. (2020). *China denies cover-up as Wuhan Coronavirus deaths revised up 50%*. Published Online. (Online news article.)
- Lardieri, A. (2020). *New York begins construction on more temporary hospitals as Coronavirus spreads*. Published Online. (Online news article.)
- Lee, H., Andrew, M., Gebremariam, A., Lumeng, J. C., & Lee, J. M. (2014). Longitudinal associations between poverty and obesity from birth through adolescence. *American Journal of Public Health*, 104(5), e70–e76.
- Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., . . . Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel Coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine*, 9(2), 538.
- Lipsitch, M. (2020). Comment: Estimating case fatality rates of COVID-19. *The Lancet Infectious Diseases*. (To appear.)
- Lipsitch, M., Donnelly, C. A., Fraser, C., Blake, I. M., Cori, A., Dorigatti, I., . . . others (2015). Potential biases in estimating absolute and relative case-fatality risks during outbreaks. *PLoS Neglected Tropical Diseases*, 9(7).



- Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS Coronavirus. *Journal of Travel Medicine*, 27(2), taaa021.
- Mizumoto, K., & Chowell, G. (2020). Early release-estimating risk for death from 2019 novel Coronavirus disease, China, January–February 2020. *Emerging Infectious Diseases*. (Early release, preprint.)
- Panackal, A. A., M'ikanatha, N. M., Tsui, F.-C., McMahon, J., Wagner, M. M., Dixon, B. W., . . . others (2002). Automatic electronic laboratory-based reporting of notifiable infectious diseases. *Emerging Infectious Diseases*, 8(7), 685.
- Paulo, A. C., Correia-Neves, M., Domingos, T., Murta, A. G., & Pedrosa, J. (2010). Influenza infectious dose may explain the high mortality of the second and third wave of 1918–1919 influenza pandemic. *PLoS One*, 5(7), e11655.
- Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A., & Jewell, C. P. (2020). Novel Coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *medRxiv*.
- Reich, N. G., Lessler, J., Cummings, D. A., & Brookmeyer, R. (2012). Estimating absolute and relative case fatality ratios from infectious disease surveillance data. *Biometrics*, 68(2), 598–606.
- Riffe, T. (2020). *COVID-19 cases and deaths by age and sex*. Published Online. (Online GitHub repository.)
- Ronco, C., & Reis, T. (2020). Kidney involvement in COVID-19 and rationale for extracorporeal therapies. *Nature Reviews Nephrology*. (To appear.)
- Sharfstein, J. M., Becker, S. J., & Mello, M. M. (2020). Diagnostic testing for the novel Coronavirus. *Journal of the American Medical Association*, 323(15), 1437-1438.
- Shi, S., Qin, M., Shen, B., Cai, Y., Liu, T., Yang, F., . . . others (2020). Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. *Journal of the American Medical Association Cardiology*. (To appear.)
- Siddique, A. (1994). Cholera epidemic among Rwandan refugees: experience of ICDDR, b in Goma, Zaire. *Glimpse (Dhaka, Bangladesh)*, 16(5), 3.
- Sliwa, K., Wilkinson, D., Hansen, C., Ntyintyane, L., Tibazarwa, K., Becker, A., & Stewart, S. (2008). Spectrum of heart disease and risk factors in a black urban population in South Africa (the Heart of Soweto Study): a cohort study. *The Lancet*, 371(9616), 915–922.
- Stelter, B. (2020). *Trump makes spurious claims about Coronavirus in phone call with Sean Hannity*. Cable News Network. Retrieved from <https://www.cnn.com/2020/03/05/media/donald-trump-sean-hannity-coronavirus/index.html> (Online news article.)
- The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. (2020). The epidemiological characteristics of an outbreak of 2019 novel Coronavirus diseases (COVID-19) in China. *China Centers for Disease Control Weekly*, 41(2), 145.

- United States Federal Government. (2020). *Remarks by President Trump, Vice President Pence, and members of the Coronavirus Task Force in press briefing on March 21, 2020*. Published Online. (Transcript published online.)
- Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., ... others (2020). Estimates of the severity of Coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases*. (To appear.)
- Wang, J., Zhu, E., & Umlauf, T. (2020). *How China built two Coronavirus hospitals in just over a week*. Published Online. (Online news article.)
- Wang, W., Tang, J., & Wei, F. (2020). Updated understanding of the outbreak of 2019 novel Coronavirus (2019-nCoV) in Wuhan, China. *Journal of Medical Virology*, 92, 441-447.
- Warne, D. J., Ebert, A., Drovandi, C., Mira, A., & Mengersen, K. (2020). Hindsight is 2020 vision: Characterisation of the global response to the covid-19 pandemic. *medRxiv*. Retrieved from <https://www.medrxiv.org/content/early/2020/05/05/2020.04.30.20085662> doi: 10.1101/2020.04.30.20085662
- Wilson, N., Kvalsvig, A., Barnard, L. T., & Baker, M. G. (2020). Early release-case-fatality risk estimates for COVID-19 calculated by using a lag time for fatality. *Emerging Infectious Diseases*. (Early release, preprint.)
- Woodruff, B., Bornemisza, O., Checchi, F., & Sondorp, E. (2014). *The use of epidemiological tools in conflict-affected populations: Open-access educational resources for policy-makers*. (Online resource.)
- Worldometer. (2020). *How to interpret the 15,152 surge in COVID-19 new cases of February 12*. (Online news article with reference to National Health Commission of China.)
- Wu, J., McCann, A., Katz, J., & Peltier, E. (2020). *28,000 missing deaths: Tracking the true toll of the Coronavirus crisis*. Published Online. (Online news article.)
- Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the Coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72, 314 cases from the Chinese Center for Disease Control and Prevention. *Journal of the American Medical Association*. Retrieved from <https://doi.org/10.1001/jama.2020.2648> doi: 10.1001/jama.2020.2648