Knowing the infection fatality rate (IFR) of SARS-CoV and SARS-CoV-2 infections is essential for the fight against the COVID-19 pandemic. A substantial amount of uncertainty in projecting the effects of the pandemic at the population level—and the impact of public policies and directives, such as social distancing measures and the impact of potential future shortages of health care supply—pivots around the uncertainty of this parameter. This parameter, as its name suggests, is the ratio of two numbers—the number of deaths caused by COVID-19 (numerator) and the total number of people in the population who were genuinely infected by the virus (denominator). However, for many reasons, both the numerator and the denominator of the IFR are measured with error. For example, errors in the denominator arise because patients remain asymptomatic during the first few days of the infection, testing is not universal and selective at best, and longitudinal data on COVID-19 patients are unavailable at the national level. Measurement errors may also exist in the numerator due to under-counting of deaths due to social isolation and other factors, and attributing some COVID-19-related deaths to other factors. Consequently, the case fatality rate (CFR) for COVID-19, which is an estimate based on the reported number of COVID-19-related deaths and the reported number of cases that were laboratory-confirmed with COVID-19 infections, provides a biased estimate of the IFR. It can be biased upwards because we do not know the actual number of individuals who are infected. It can be estimated at the county level.

Using data through April 20, 2020, we fit a statistical model to COVID-19 case fatality rates over time at the US county level to estimate the COVID-19 IFR among symptomatic cases (IFR-S) as time goes to infinity. The IFR-S in the US was estimated to be 1.3% (95% central credible interval: 0.6% to 2.1%). County-specific rates varied from 0.5% to 3.6%. The overall IFR for COVID-19 should be lower when we account for cases that remain and recover without symptoms. When used with other estimating approaches, our model and our estimates can help disease and policy modelers to obtain more accurate predictions for the epidemiology of the disease and the impact of alternative policy levers to contain this pandemic. The model could also be used with future epidemics to get an early sense of the magnitude of symptomatic infection at the population level before more direct estimates are available. Substantial variation across patient demographics likely exists and should be the focus of future studies. [Editor’s Note: This Fast Track Ahead Of Print article is the accepted version of the peer-reviewed manuscript. The final edited version will appear in an upcoming issue of Health Affairs.]
biased downward because some of those who are currently infected could die in the future, or deaths are undercounted. The upward bias is likely to be much larger during the early phase of testing. Most estimates of the COVID-19 fatality rate currently available around the world suffer from these biases.

In this paper, we try to overcome these biases using national US data on counts of reported deaths and detected COVID-19 cases and the temporality of CFR (that is, its variability over time) to make inferences about the IFR for COVID-19. Our method does not account for a fraction of cases with the COVID-19 infection that recovers without any major symptoms. These asymptomatic patients do not contribute to any of the reported statistics on COVID-19 deaths and cases. A true IFR should include these patients in the denominator. However, in the paper, since we try to eliminate measurement errors in reported CFRs based on trends in reported COVID-19 deaths and cases, we are unable to account for this fraction of the population who remain asymptomatic with infections. Consequently, what we estimate is the IFR among symptomatic COVID-19 cases (IFR-S).

**Study Data And Methods**

**Assumptions** We make three assumptions for our analysis: (1) Errors in the numerator and the denominator lead to underreporting of true COVID-19 deaths and cases, respectively; error is smaller for deaths than for cases. (2) Both the errors are declining over time. (3) The errors in the denominator are declining at a faster rate than the error in the numerator. Assumption #1 is self-evident; both the deaths and the actual cases are undercounted during the initial phase of the epidemic.\(^3,4\) Because deaths are much more visible events than infections, which, in the case of COVID-19, can go asymptomatic during the first few days of infection, we posit that, at any point in time, the errors in the denominator are larger than the errors in the numerator. Hence, this assumption leads to CFR estimates being larger than the IFR-S, which is typically believed to be true based on observed data.

Assumption #2 is our central assumption, which states that under some stationary processes of care delivery, health care supply, and reporting, which are all believed to be improving over time, the errors in both the numerator and the denominator are declining. It implies that we are improving in the measurement of both the numerator and denominator over time, albeit at different rates in different jurisdictions.

Assumption #3 posits that the error in the denominator is declining faster than the error in the numerator. This assumption indicates that the CFR rates, based on the number of cumulative COVID-19 deaths and the cumulative reported COVID-19 cases, are declining over time and are confirmed based on our observed data (described in detail below).

If these simple assumptions hold, our methods allow us to project the IFR at the limit when time goes to infinity and errors reduce to zero. That is not to say that, in practicality, we expect that in the future, we would ever reach a point where we would have universal testing or comprehensive reporting of all COVID-19 deaths. However, as long as we are improving and reducing the errors, we can make inferences about the IFR-S by fitting models on the temporality of CFR and thereby projecting the expected rate at infinite time.

However, this stationary process of declining errors may be disrupted in certain regions, and over certain time-periods, by shortages of testing supplies, leading to artificial increases in CFR after days of decline. We apply specific criteria to identify these regions and time-periods so that we can exclude these data from our analysis.

We provide a detailed mathematical formulation of these assumptions and their implications in the online appendix.\(^5\) We use an exponential decay formulation within a logit framework to model this decay to estimate the asymptote parameter. In plainer language, the fact that reported case fatality rates are observed to decline over time early in an epidemic, we are able to use statistical techniques to estimate the value of those rates as they approach their “limit” when the decline continues over a long time horizon. This rate at the “limit” presumably reflects the value of the IFR-S. More importantly, we allow for heterogeneity in decay across US counties to estimate our target parameter. Our estimate of IFR-S is less sensitive to the missing deaths that will occur in the future from the cases detected in the last two weeks of the limit. We discuss this implication in the Conclusion Section. We validate our predictions based on declining observed rates during the future dates that were not used for estimation purposes.

**Data** We used publicly reported data in GitHub from the Johns Hopkins Repository\(^6\) and the New York Times\(^3\) on the total number of cumulative deaths and detected cases by day for each county in the United States. We updated missing values from one repository using the non-missing values from the other by date and county. Moreover, for any date and county, the maximal value reported for deaths or detected cases in either repository was used. A rate variable was constructed by dividing the cumulative total
number of deaths by the cumulative total detected cases for each date and county. The first diagnosed case of COVID-19 in the US occurred on January 21 in Washington State, while the first death, also in Washington State, occurred on February 28, although new data are showing that earlier cases may have existed in California. Since testing was non-existent during the initial few days, the data showed that this ratio increased for the first few days for many counties. Therefore, for each county, our analysis started from the day when the first zenith in this rate was reached. It is assumed that the declining error rates within each county began from that day forward driven by better reporting of COVID-19 deaths and cases. Only counties that had reported at least five COVID-19 deaths and thirty cases up to April 20, 2020, were retained.

Moreover, we were aware that sudden area-wide shortages in testing kits could artificially raise the CFR after days of decline, and, therefore, bias our decay analysis. That is why we removed counties that reported at least one standard deviation increase in CFR for seven or more days after reaching the CFR nadir. Also, among the remaining counties, we removed the last seven days of follow-up if CFRs were found to increase consecutively for three or more days during that week. Lastly, we kept counties that had at least six follow-up days of reported data after reaching the zenith.

**Statistical Model** We modeled these rates over time for each county using a Binomial model for the counts of deaths over the counts of detected cases, i.e., \( \text{Deaths}_j \), are distributed as Binomial\((p_{jt}, \text{Detected}_j) \), where \( j \) denotes the counties and \( t \) denotes the number of days from the zenith value of rate within a county \( \text{(Days)} \). The mean of this Binomial model, \( p_{jt} \), represents the probability of death, and is expressed as a Bayesian random-coefficients exponential decay model within a logit link framework so that the predicted rates remain within 0 and 1. Specifically, our model estimated the probability of death in county \( j \) at time \( t \) equals Logit\( ^{-1}(A_{1j} + \exp(-\exp(A_{3j})(\text{Days}_j - 1))) \).

The main feature of the decay model is that as time \( \text{(Days)} \) goes to infinity, under the assumption that errors in both the numerator and denominator go to zero, the cumulative CFR would approach an estimate of the true IFR-S in the population. Specifically, in this model:

\[
\text{Logit}^{-1}(A_{1j}) = \text{County-specific IFR-S, as Days goes to infinity.}
\]

\[
\text{Logit}^{-1}(A_{2j}) = \text{County-specific exponential decline rate in CFR, parameterized such that it takes on negative values only.}
\]

The overall US-specific IFR-S can be expressed as Logit\( ^{-1}(b_1) \), where \( A_{1j} \) is distributed as Normal \( (b_1, 1) \). Hyperpriors for coefficients were based on Cauchy distributions, as recommended in the Bayesian literature for logistic models.

Prior sensitivity analyses were carried out based on using Normal or Uniform distribution for the hyperpriors. Further details about the model are available in the appendix. We use the Metropolis-Hastings algorithm to estimate this model, using three simultaneous Monte-Carlo Chains and 10,000 deviates for each chain, 10,000 burn-in runs, and a thinning of 100.

We used data up to April 20, 2020 for our training sample to estimate our model. Model fit was assessed using posterior predictions from the model against four consecutive follow-up days for each county. For most counties, these days were April 21–24.

**Limitations** There are several limitations to our analysis. First, we acknowledge that our estimate of IFR-S would be higher than the true overall IFR. This is because our model relies on identified cases who are presumably all symptomatic COVID-19 patients. Therefore, even at the limit, our estimated rate would not include the fraction of patients who may have the infection but remain and recover asymptptomatically. Our estimate, however, would include patients who start with an asymptomatic infection but become symptomatic later on. An estimate of the magnitude of the truly asymptomatic fraction in COVID-19 remains unclear—population-wide antibody testing would be needed to establish this statistic. Results from sero-testing from the Diamond Princess outbreak suggests that about 17.9% of infected persons never developed symptoms. Consequently, a reasonable estimate of the overall IFR would be about 20% lower than our estimated IFR-S.

Second, our estimated COVID-19 IFR-S may be slightly conservative. Our approach was to control for the upward bias in this estimate if raw rates were used. We do not control for the downward bias that may arise because some of the detected cases may die in the future. Recently, there was an attempt to address this downward bias by estimating lagged death rates based on international data and estimated an effect ranging from 0.8% (China excluding Hubei province) to 4.2% in 82 other countries and territories. They used a time lag of 13 days, based on reported data from China on time from radiologic confirmation of COVID-19 to death. However, since the distribution of time-to-death varies over this 13-day follow-up period, such a correction could give non-sensical results when applied to US data that has less than 60 days of COVID-19 history. The death rate was estimated.
to be above one on specific days for many US counties when such a correction was applied. In general, we believe that the downward bias generated due to the missing deaths at the limit should be small. This is because our estimate represents the death rate in an asymptote, which is what would happen with many days of accumulated data on the number of detected and death cases. At that point, the additional amount of deaths from that last two weeks of detected cases would contribute very little to our overall estimate of the IFR-S.

Third, what we present here are crude IFR-S and not even age-adjusted ones. We did not have any data to assess the distribution of IFR-S across age and comorbidity profiles of patients. One would need, ideally, individual-level data, and at the least group-specific data to estimate such dispersion, which are not publicly available.\textsuperscript{13,14} The Centers for Disease Control and Prevention reports a significant variation in fatality rates by age groups.\textsuperscript{15} Further work is required on this front.

### Study Results

One thousand three hundred and sixty-four counties out of 3020 counties in the United States reported any confirmed COVID-19 case by April 20, 2020. Of these, 134 counties reported no COVID-19 deaths until that time. One thousand thirty-four counties had any reported COVID-19 deaths by April 20. Three-hundred ninety-seven counties reported exactly one COVID death by this date. By April 20, there were 753,113 confirmed cases of COVID-19, and 41,287 reported COVID-19 deaths. For our analysis, we used 116 counties. Interestingly, New York County, NY, (FIPS: 36061; does not represent the entire New York City) was not included in our analysis, despite having the highest number of cases and deaths in the country. The number of deaths in this county was rising at a faster rate than the number of detected until April 20, 2020; hence, the CFR was yet to reach a zenith. Overall, a total of 40,835 confirmed cases and 1,620 confirmed deaths until April 20 were used for our analysis.

The 116 counties selected spanned 33 states, with Georgia contributing the maximum with thirteen counties, followed by Louisiana with nine, and then by South Carolina with eight. After reaching the initial zenith, CFRs were found to be declining within each of these retained counties, supporting our assumptions about the differential declining error rate between the numerator and denominator of the CFRs for these counties. A description of growth in COVID-19 reported cases and deaths, and the decline in the CFRs is available in the appendix.\textsuperscript{5} At the zenith of the computed rate in each county, the rate variable varied from 1.7% to 33.3%. By the end of follow-up, the rate varied from 0.9% to 19.3% across counties. Follow-up days ranged from 7 to 31.

The Bayesian model showed good convergence and mixing properties between the model and the observations. Gelman-Rubin statistics were below 1 for each of the parameters of the model, indicating that the three independent Monte-Carlo chains overlapped and converged to similar posterior distributions for the parameters. These results are available in the appendix,\textsuperscript{5} including residual analysis based on fitted posterior means (means predicted by the model for the period prior to the validation phase) from our prediction model, which appears to fit the county-level data over time well. The posterior mean of the US-specific IFR-S was estimated to be 1.3% (median: 1.3%, Std. Dev: 0.4%) with a 95% central credible interval of 0.6% to 2.1% (see supplemental exhibit 1).\textsuperscript{5}

The posterior means and the 95% credible intervals of county-specific IFR-Ss for twenty counties with the lowest rates (0.5%–1.4%) (plus the overall values for the U.S.) and twenty-one counties with the highest rates (2.3%–3.6%) are shown in supplemental exhibits 1 and 2, respectively.\textsuperscript{3} The IFR-S for other counties in the middle that are not shown ranged from 1.5% to 2.2% and are described in the appendix.\textsuperscript{5} The lowest rate was estimated to be in Putnam County, NY, (0.5%, 95% CCI: 0.1%–1.0%) while the highest in King County, WA (3.6%, 95%CCI: 0.5%–6.1%). Data at the county level are still evolving, and hence considerable uncertainty exists for some counties, especially towards the higher range of IFR-S estimates. Since these estimates represent the crude IFR-S, many factors contribute to their variation across counties, including demographics (especially age distribution), levels of population health, and supply of healthcare. In that sense, IFR-S is a dynamic quantity even within a county, depending on how the case-mix of the infected population shifts over time.

To assess the validity of our prediction model, we forecasted county-specific rates based on the posterior predictive mean over four days following the estimation time window for each county and compared those with observed rates during these days.\textsuperscript{16} Supplemental exhibit 3\textsuperscript{5} presents the comparison between predicted and observed rates, where each dot in the exhibit represents a rate for a county on a given day. It shows that the 95% central credible intervals from the posterior predictive distribution from the model were able to capture the true CFR rates (represented by the
45-degree diagonal line) for all counties over these four days. The Bayesian posterior predictive two-sided p-values\textsuperscript{17,18} were less than 0.05 for none of the 116 counties for any of the four days.

**Conclusion**

After modeling the available national data on cumulative deaths and detected COVID-19 cases in the United States, the IFR-S from COVID-19 was estimated to be 1.3%. This estimated rate is substantially higher than the approximate IFR-S of seasonal influenza, which is about 0.1\%\textsuperscript{19} (34,200 deaths among 35.5 million patients who got sick with influenza). Influenza is also believed to be completely asymptomatic in 16\% of the infected population,\textsuperscript{20} and this fraction is not included in the calculation of its IFR-S.\textsuperscript{21} Our COVID-19 IFR-S estimate is not outside the ballpark of estimates becoming available from other countries, but certainly lower, as expected from addressing the upward bias in those estimates. For example, the COVID-19 fatality rate for China (without correction for the upward bias inherent in looking at observed rates) was initially reported to be 5.6\% (95\% CI: 5.4–5.8\%).\textsuperscript{22} By February 20, the crude fatality rate for China was estimated to be 3.8\%.\textsuperscript{23} The fatality rate outside China was estimated to be 15.2\% (95\% CI 12.5–17.9\%),\textsuperscript{22} which may be due to the more considerable upward bias during the beginning part of the pandemic within a county. We see the same patterns in the United States, with observed rates being much higher during the initial part of the pandemic. A recent estimate of CFR using individual-level data from Wuhan residents and also international Wuhan residents who repatriated on six flights was found to range from 0.66\% to 1.4\%.\textsuperscript{24}

If we carry out a thought experiment where 35.5 million individuals would contract COVID-19 illness this year in the US (i.e., the same number as flu last year)\textsuperscript{19} then, in the absence of any mitigation strategies or social distancing behaviors and the supply of health care services under typical conditions, our IFR-S estimate predicts that there would have been nearly 500,000 COVID-19 deaths this year. To the extent that COVID-19 is more infectious than flu and does not have any protection from a vaccine or treatment, the number of infections, and hence the number of deaths, would be higher. Certainly, with mitigation strategies, the death toll will be lower. For example, the recent White House COVID-19 Taskforce projections of 100,000–200,000 deaths this year from COVID-19 is made with assumptions about the effectiveness of social distancing directives and measures currently in place.\textsuperscript{25}

Our estimated IFR-S applies under the assumption that the current supply (until April 20) of health care services, including hospital beds, ventilators, and access to healthcare providers, would continue in the future. Constraints in the supply of health care services could surely increase IFR and the overall fatality rates. We hope that simulations to understand and forecast the impact of such shortages can be improved using our estimates of IFR-S as the baseline.

Similarly, our estimates of the COVID-19 IFR-S in the US can help disease and policy modelers to obtain more accurate predictions for the epidemiology of the disease and the impact of alternative policy levers to contain this pandemic.

The author received compensation from Salutis Consulting LLC. No funding was received for this analysis. The author thanks Varun Gandhay for excellent research assistance for this work. He also thanks six anonymous reviewers and Donald Metz for their excellent comments. The views expressed do not represent those of the University of Washington or the NBER. All errors are those of the author [Published online May 7, 2020].

NOTES


5 To access the appendix, click on the Details tab of the article online.

6 Johns Hopkins University Covid-19 Data Repository [Internet]. Baltimore (MD): Johns Hopkins
COVID-19