Supplementary Information for "Estimating geographic variation of infection fatality ratios during epidemics"

Joshua Ladau*[1,2,3], Eoin L. Brodie[4], Nicola Falco[4], Ishan Bansal[3], Elijah B. Hoffman[2,5], Marcin P. Joachimiak[6], Ana M. Mora[7], Angelica M. Walker[8], Haruko M. Wainwright[9], Yulun Wu[5], Mirko Pavicic[10], Daniel Jacobson†[10], Matthias Hess†[11], James B. Brown†[2,3,12], and Katrina Abuabara[1,13]

[1]*Departments of Computational Precision Health and Dermatology, University of California, San Francisco, CA 94115*

[2]*Arva Intelligence, Inc., Salt Lake City, UT 84101*

[3]*Computational Biosciences Group, Lawrence Berkeley National Laboratory, Berkeley, CA 94720*

[4]*Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720*

[5]*Graduate Group in Biostatistics, University of California, Berkeley, CA 94720*

[6]*Biosystems Data Science, Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720*

[7]*Center for Environmental Research and Community Health (CERCH), School of Public Health, University of California, Berkeley, CA 94720*

[8]*Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN 37996*

[9]*Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Boston, MA 02139*

[10]*Biosciences, Oak Ridge National Laboratory, Oak Ridge, TN 37830*

[11]*University of California, Davis, CA 95616*

[12]*Statistics Department, University of California, Berkeley, CA 94720*

[13]*Division of Epidemiology and Biostatistics, University of California Berkeley School of Public Health, 2121 Berkeley Way, Berkeley, CA 94720*

*Corresponding author: jladau@gmail.com.

†Authors contributed equally

1 Supplementary Methods

1.1 Point Estimate Calculation

To apply the estimators to the COVID-19 pandemic, we calculated estimates in the following sequence:

1. T_c: Using Assumption (iv) listed above, for each week interpolate the number of SARS-CoV-2 NAAT tests performed in each county where testing data is unavailable.

2. T_s and T: For each week, use the sums of the numbers of tests in each county to find the number of tests in each state and across the entire U.S.

3. ι: For each week, estimate the total number of SARS-CoV-2 infections across the U.S. using T from the previous step, the estimator $\hat{\iota}$, and Assumption (ii).

4. ω: For each week, estimate the odds ratio across the U.S. using the estimator $\hat{\omega}$.

5. ω_s and ω_c: For each week, estimate the odds ratio at the state and county levels via the Assumption (i) listed above.

6. ι_s and ι_c: For each week, estimate the number of SARS-CoV-2 infections at the state- and county-levels using the estimator $\hat{\iota}_j$ and the odds ratios from the previous step. Sum these estimates across weeks to generate estimates of the total number of SARS-CoV-2 infections between April 1, 2020 - September 30, 2020.

7. ϕ_s and ϕ_c: Estimate the SARS-CoV-2 IFRs at the state- and county-levels using the estimator $\hat{\phi}_j$ with the total mortality and estimated infections between April 1, 2020 - September 30, 2020.

1.2 Validation: performance when assumptions are met

To assess the performance of the estimators, under scenarios with set numbers of infections, we generated simulated COVID-19 case data for each week between April 1, 2020 to September 30, 2020, and then numerically assessed the performance of the estimators against these known numbers of infections. Specifically, we used the following generating model:

1. We set the country-level odds ratio, ω_n equal to $26/\sqrt{d+0.5}$, where d is the number of weeks elapsed since April 1, 2020. (With regard to the numerator, there were 26

weeks between April 1, 2020 and September 30, 2020.) As per Assumption (i), we set the state- and county-level odds ratios equal to the same country-level values.

2. For each region, with the exception of the numbers of cases and country-wide IFR, we set the observed quantities in Table S1 equal to their known values (Table S2).

3. For each county, we simulated the total number of SARS-CoV-2 infections (ι_c) as a random variate from the following distribution: $\lfloor 0.0196 \cdot P_c U^4 \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function and U is a uniform $[0, 1]$ random variate, with the condition that $\iota_c \geq D_c$. While this may seem like an arbitrary choice of distributions, it has the desirable property of resulting in a country-wide IFR (ϕ_n) of 5.00 deaths per $1,000$ infections, consistent with Assumption 2.

4. For each county, we simulated the number of COVID-19 cases in each county (C_c) as a variate from Wallenius' noncentral hypergeometric distribution, using the simulated odds ratio and number of SARS-CoV-2 infections from above.

5. For each state and the entire country, we found simulated numbers of SARS-CoV-2 infections and COVID-19 cases (ι_s, ι_n, C_s, and C_n) by summing the corresponding values simulated above from the counties that it contained.

2 Supplementary Figures

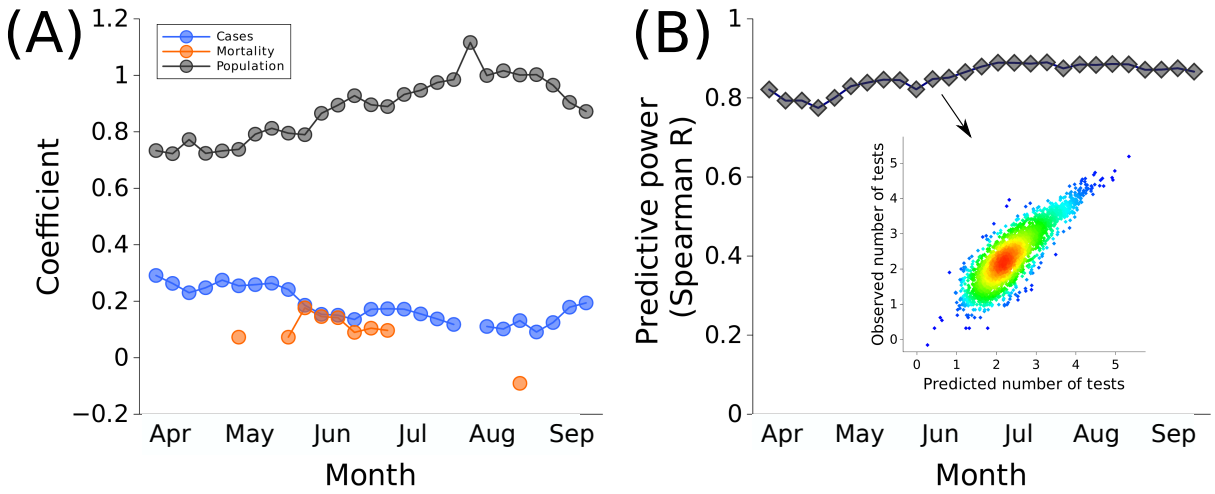

Figure S1: Performance of models used to predict the number of SARS-CoV-2 NAATs in counties where testing data were unavailable. Separate linear models were fit for each week between April 1, 2020 and September 30, 2020, with the total population and number of observed COVID-19 cases and mortality used as possible predictors, and the number SARS-CoV-2 NAATs as the response (all log transformed). (A) All subsets model selection with 51-fold cross validation yielded population and the number of COVID-19 cases as consistently the best predictors. Missing symbols indicate that a predictor was not included in the model for a given week. (B) The models generally had high predictive accuracy, with the correlation between predicted and observed (omitted in cross-validation) values consistently greater than 0.8. The inset graph shows an example of the predictive power for the second week of June. Each point represents the number of SARS-CoV-2 NAATs in a held-out county; color indicates the density of points.

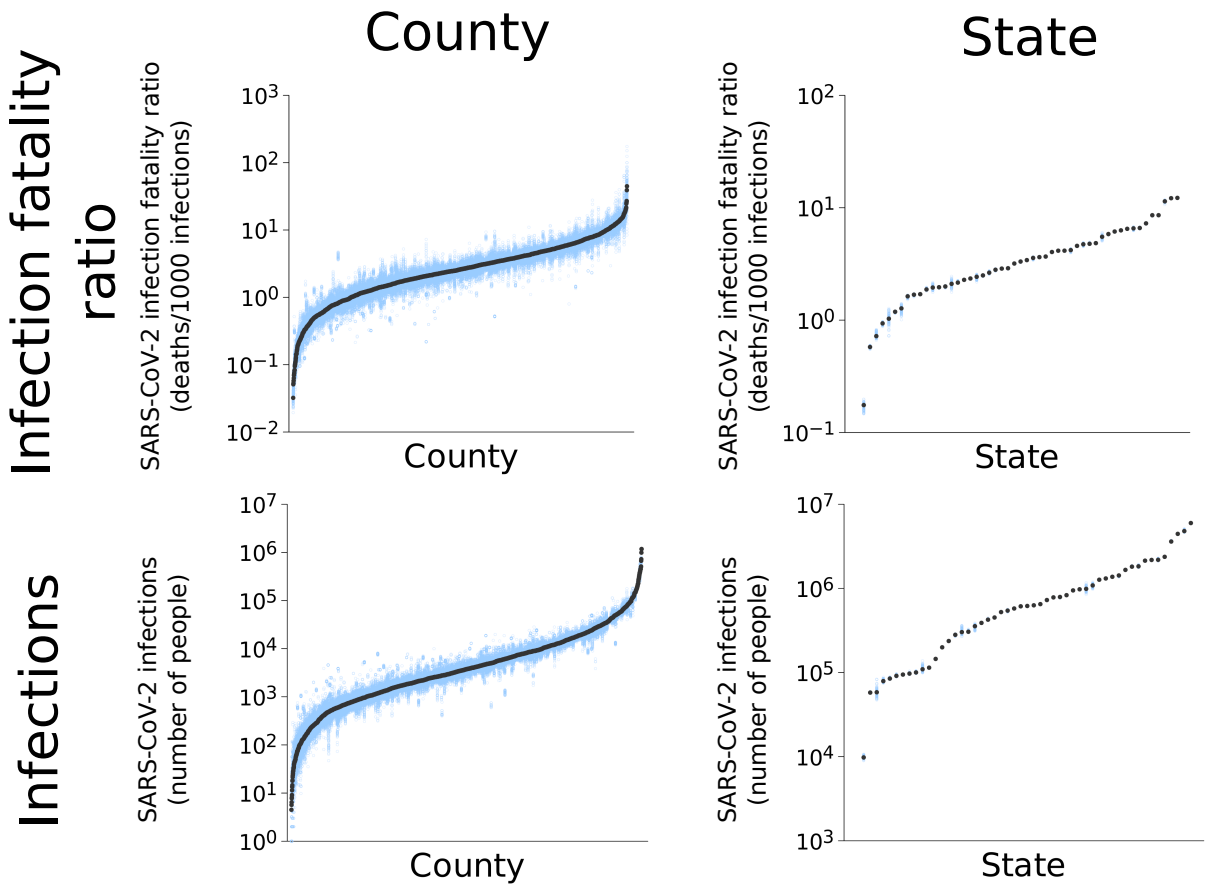

Figure S2: Uncertainty associated with the point estimates for the SARS-CoV-2 IFRs and numbers of infections. Counties and states (x-axes) are ordered by increasing point estimate. Point estimates are shown by black points; bootstrap resamples ($n = 100$ for each county and state) are shown in blue. Both the IFR and infections estimators have low variance at both the county and state level.

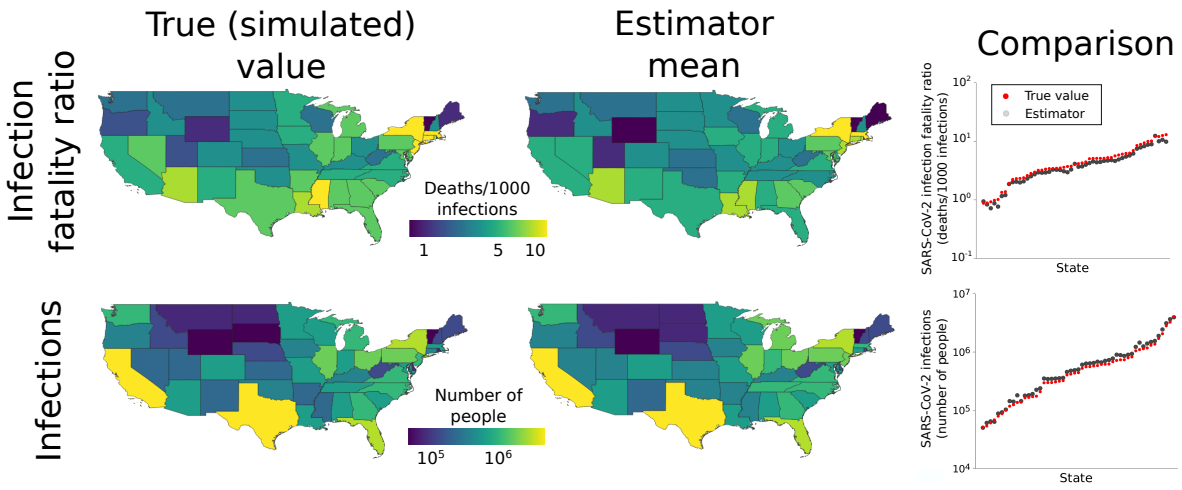

Figure S3: Performance of state-level estimators. The maps and graphs are analogous to those in main text Figure 2, but are at the state level.

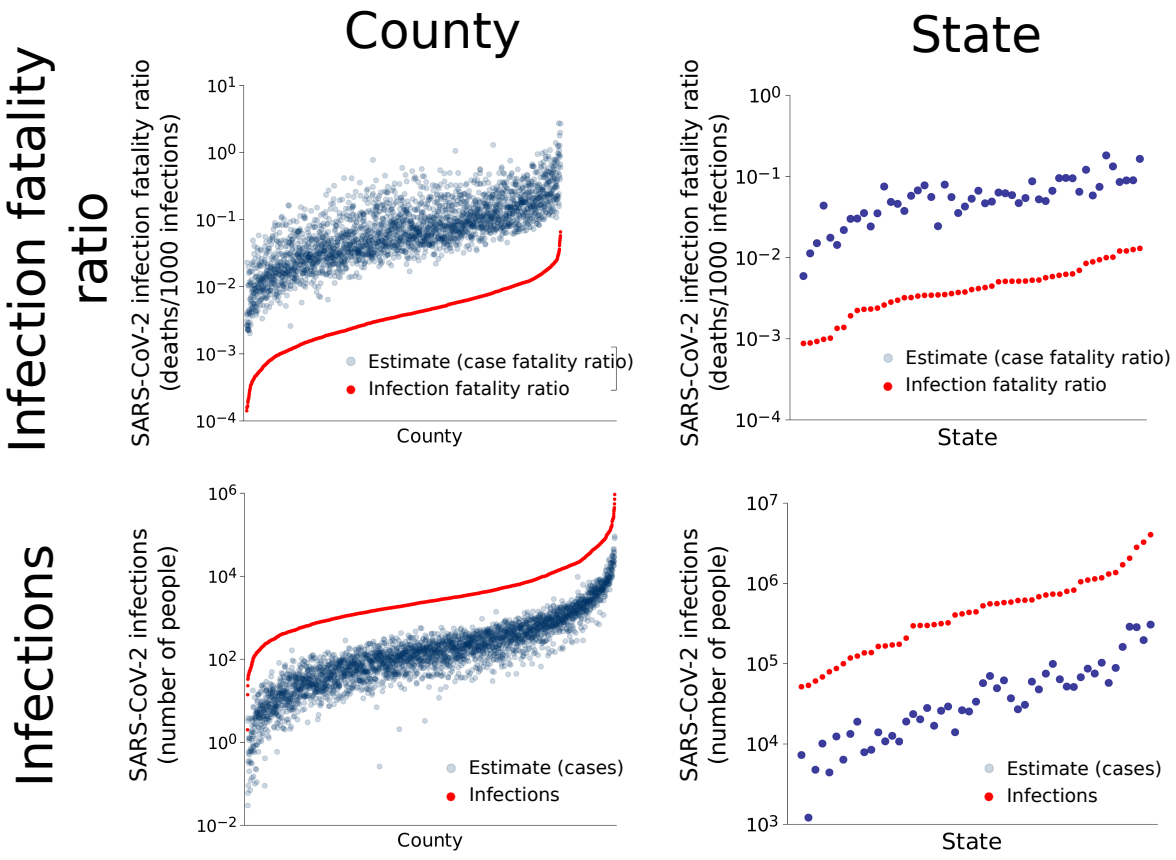

Figure S4: The estimators developed here greatly outperform uncorrected estimators for IFRs and numbers of infections. With simulated data where true values were known, at both the county and state levels, the case fatality ratio and number of cases overestimate and underestimate the IFR and number of infections, respectively, by up to several orders of magnitude. While this might seem expected, it shows that the estimators for developed here for the IFR and number of infections make substantial corrections.

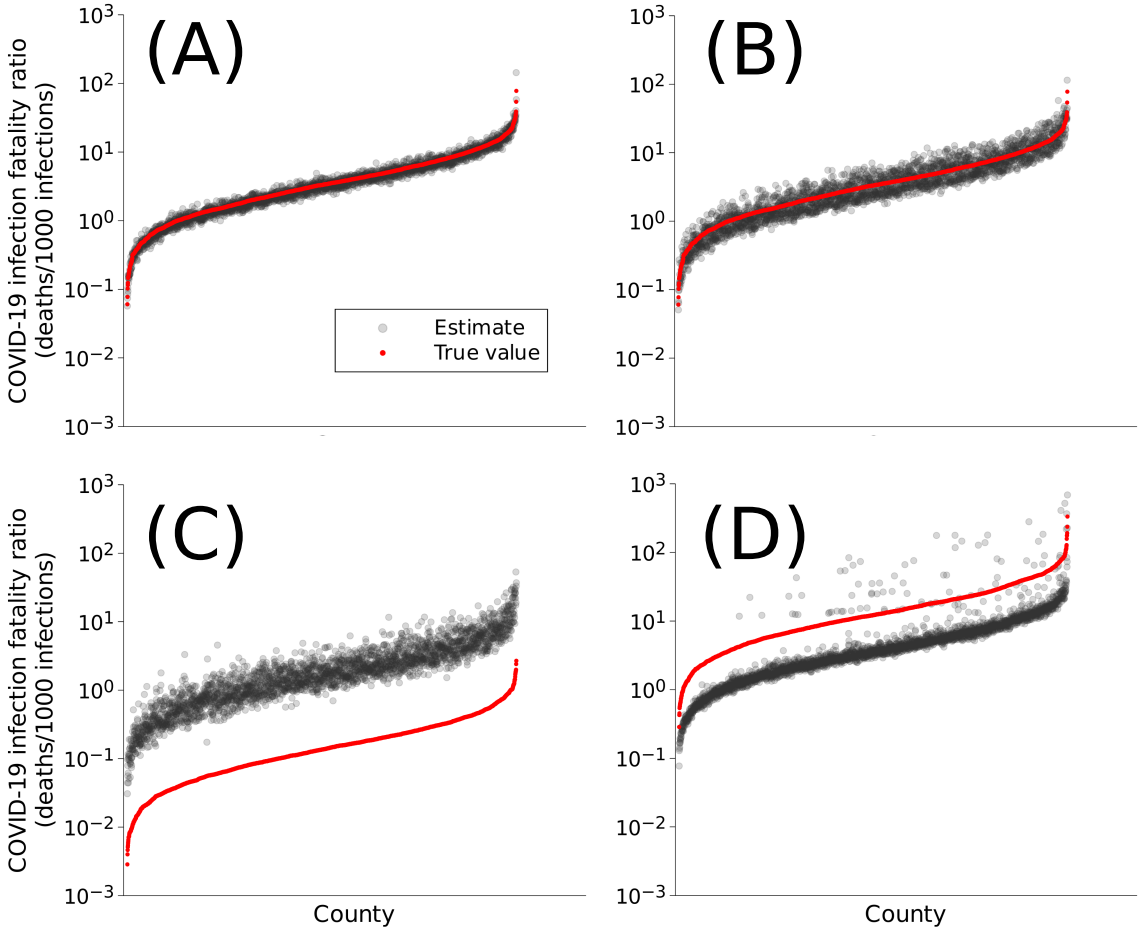

Figure S5: The estimators at the county level are relatively robust against model misspecification. (A) Low and (B) high levels of geographic variability in the odds ratios, a violation of Assumption 1, result in an increase in the variance of the estimators, but little bias. Overall IFRs (C) higher and (D) lower than the assumed value of 5 deaths per 1,000 infections lead bias of the estimators, while retaining the ability of the estimators to correctly rank IFRs in different geographic regions relative to each other.

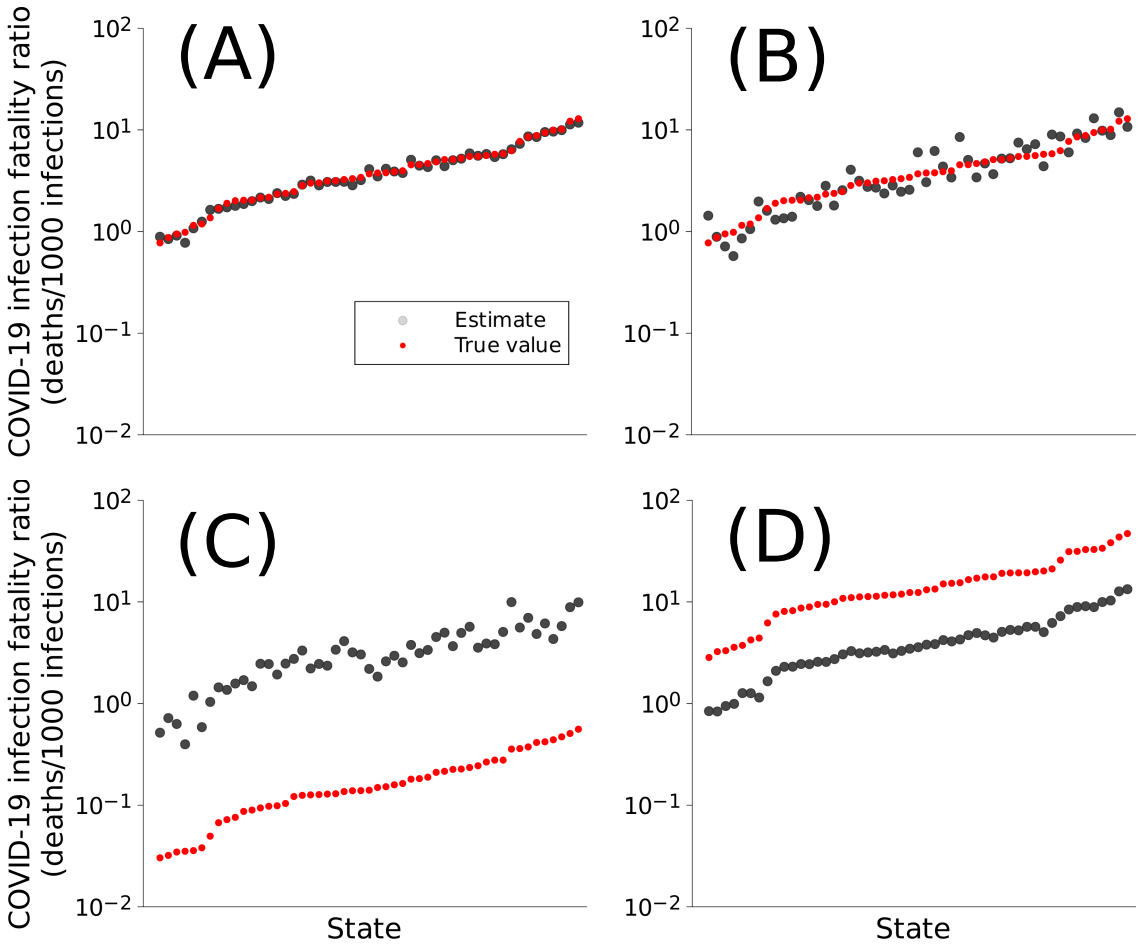

Figure S6: The estimators at the state level are relatively robust against model misspecification. As in Figure S5, there is a slight increase in variance with (A) low and (B) high levels of geographic variability in the odds ratios, a violation of Assumption 1. Bias in the overall IFRs (C) higher and (D) lower than the assumed value of 5 deaths per 1,000 leads to bias in the estimators, but they retain the ability to correctly rank geographic locations.

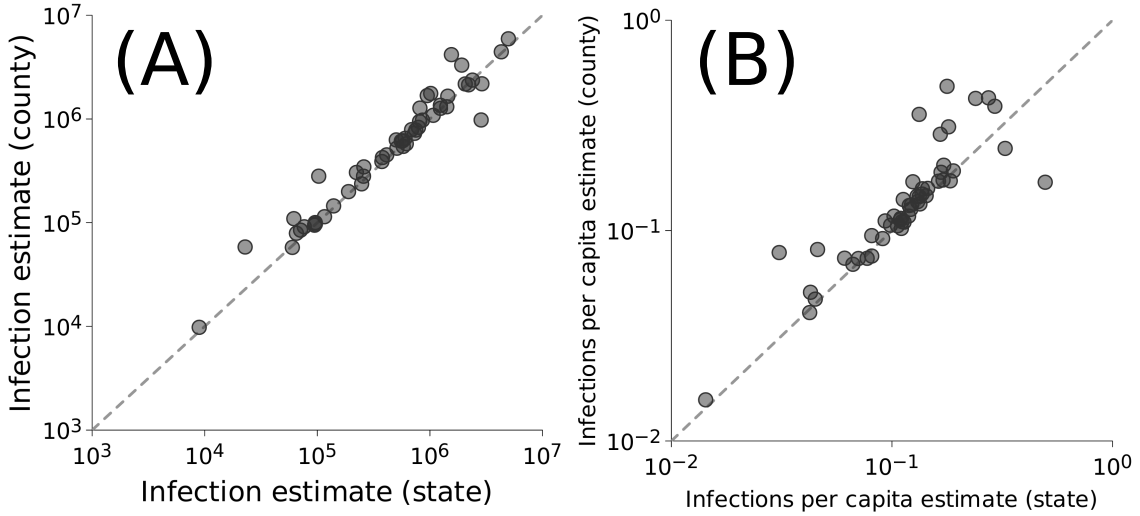

Figure S7: The estimates of the numbers of SARS-CoV-2 infections, directly estimated for states, match the numbers estimated by summing the SARS-CoV-2 infection the estimates for the counties, regardless of whether the estimates are (A) uncorrected or (B) corrected for total state population. The consistency of these estimation approaches suggests good performance of the estimators.

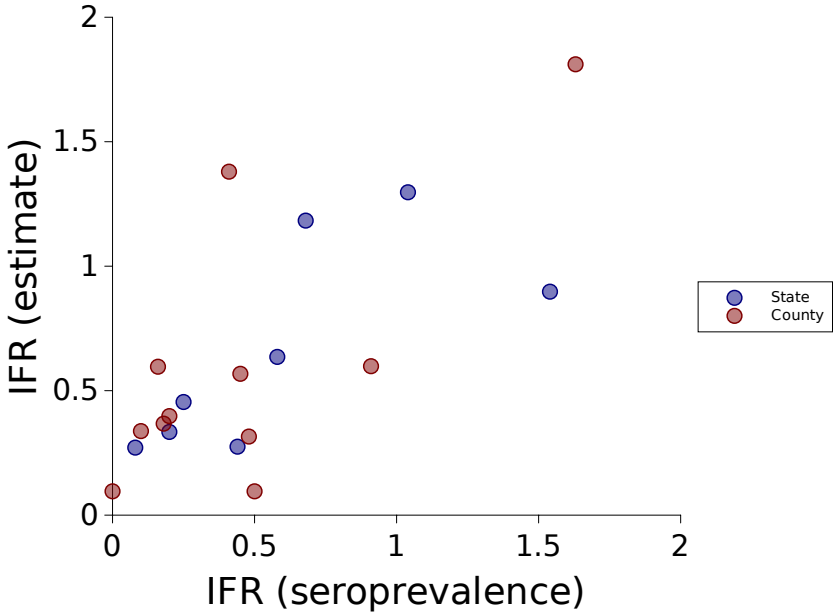

Figure S8: Comparison between SARS-CoV-2 IFR estimates from the hypergeometric estimator and independent seroprevalence estimates from other studies Ioannidis (2021). At both the state and county levels, there is a strong match between the estimate types (most points fall close to the 1:1 line).

3 Supplementary Tables

Table S1: Real world COVID-19 data: Observed (black) and unobserved (red) quantities. The number of tests was partially observed and estimated elsewhere.

Quantity	Country	State	County
Deaths	D_n	D_s	D_c
Cases	C_n	C_s	C_c
Population	P_n	P_s	P_c
Infection fatality ratio	ϕ_n	$\boldsymbol{\phi_s}$	$\boldsymbol{\phi_c}$
Infections	$\boldsymbol{\iota_n}$	$\boldsymbol{\iota_s}$	$\boldsymbol{\iota_c}$
Tests	T_n	T_s	T_c
Odds ratio	$\boldsymbol{\omega_n}$	$\boldsymbol{\omega_s}$	$\boldsymbol{\omega_c}$

Table S2: Validation data: Observed (black) and simulated (blue) quantities. Starred quantities were simulated as being non-random.

Quantity	Country	State	County
Deaths	D_n	D_s	D_c
Cases	$\boldsymbol{C_n}$	$\boldsymbol{C_s}$	$\boldsymbol{C_c}$
Population	P_n	P_s	P_c
Infection fatality ratio	$\boldsymbol{\phi_n}$	$\boldsymbol{\phi_s}$	$\boldsymbol{\phi_c}$
Infections	$\boldsymbol{\iota_n}$	$\boldsymbol{\iota_s}$	$\boldsymbol{\iota_c}$
Tests	T_n	T_s	T_c
Odds ratio	$\boldsymbol{\omega_n *}$	$\boldsymbol{\omega_s *}$	$\boldsymbol{\omega_c *}$