Bayesian estimation of SARS-CoV-2 prevalence in Indiana by random testing

Constantin T. Yiannoutsos[a,1], Paul K. Halverson[b], and Nir Menachemi[b,c]

[a]Department of Biostatistics, Indiana University Fairbanks School of Public Health, Indianapolis, IN 46202; [b]Department of Health Policy and Management, Indiana University Fairbanks School of Public Health, Indianapolis, IN 46202; and [c]Regenstrief Institute, Inc., Indianapolis, IN 46202

Edited by Adrian E. Raftery, University of Washington, Seattle, WA, and approved December 2, 2020 (received for review July 2, 2020)

From 25 to 29 April 2020, the state of Indiana undertook testing of 3,658 randomly chosen state residents for the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus, the agent causing COVID-19 disease. This was the first statewide randomized study of COVID-19 testing in the United States. Both PCR and serological tests were administered to all study participants. This paper describes statistical methods used to address nonresponse among various demographic groups and to adjust for testing errors to reduce bias in the estimates of the overall disease prevalence in Indiana. These adjustments were implemented through Bayesian methods, which incorporated all available information on disease prevalence and test performance, along with external data obtained from census of the Indiana statewide population. Both adjustments appeared to have significant impact on the unadjusted estimates, mainly due to upweighting data in study participants of non-White races and Hispanic ethnicity and anticipated false-positive and false-negative test results among both the PCR and antibody tests utilized in the study.

COVID-19 | SARS-CoV-2 | random sample

The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a viral strain that causes COVID-19. Since its emergence in late 2019, it has resulted in a pandemic of historical proportions. In the United States, there were over 1 million cases confirmed by the end of April 2020 and almost 65,000 deaths (1). However, the true extent of the pandemic is not known due to a number of contributing factors. Most importantly, a significant proportion of people infected with the virus are asymptomatic or show mild symptoms (2). In addition, due to the novelty of this virus and the speed of its spread, diagnostic test availability was limited at the time of the study, so testing was concentrated on symptomatic individuals or those with symptoms severe enough to warrant hospitalization. Recently, one study was published (3) (the University of Southern California study; henceforth, the "USC study"), and one has been posted in preprint archives (4) (the "Stanford study"), describing efforts to estimate the prevalence of infection through testing nonhospitalized individuals. These studies suggest that the majority of COVID-19 cases may go undetected, with significant implications for economic, policy, and public health decision making. Given the high-stakes nature of the pandemic, these papers have received unusual scrutiny, despite (or because of) their limited geographic coverage or concerns about the validity and generalizability of their findings due to lack of random selection of their subjects.

Generally, two types of tests are used. A molecular test, predominantly administered through nasopharyngeal swabs, which assesses the possibility of having current infection, and an antibody test that detects the presence of previous infections in blood serum. Several tests of each type have received Emergency Use Authorization by the US Food and Drug Administration (5). However, data on their accuracy are limited. Even more limited are studies assessing biases in prevalence estimation resulting from testing errors, both false-positive (where uninfected subjects test positive for the disease) or false-negative (where infected subjects test negative on the diagnostic test). Notable exceptions to this include a small study by Qian et al. (6) and Gelman and Carpenter (7), along with the sensitivity analyses undertaken by Bendavid et al. in the Stanford study (4). At the time of this writing, this latter study is in the peer-review stage.

This report describes the statistical analysis of SARS-CoV-2 testing data in the state of Indiana (henceforth, the "Indiana study"). The Indiana Department of Health (IDOH) undertook statewide testing between 25 and 29 April 2020, involving randomly chosen residents from the state. At the time, this was the only statewide cohort of randomly selected individuals to be tested for infection with the SARS-CoV-2 in the United States (8). As of this writing, the Indiana study and two subsequently completed waves of random testing of Indiana state residents are the only randomized statewide studies for COVID-19 ever performed in the United States. While study participants were selected randomly, however, substantial nonresponse and concerns about diagnostic testing errors raise the possibility of significant biases present in the unadjusted estimates of disease prevalence. The present paper describes methods used to address a number of these concerns.

Data Sources

Data used in this study were obtained from three sources: 1) census data for all counties in Indiana, with summaries by sex, age, race and ethnicity, along with margins of error for these estimates; 2) the results from the Indiana statewide testing, a sample selected according to a stratified random-sampling design, performed between 25 and 29 April 2020; and 3) information on the number of daily confirmed COVID-19 cases and deaths resulting from COVID-19 disease, provided by the IDOH dashboard (10),

Significance

Infection with the novel coronovirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in a worldwide pandemic of COVID-19 disease. Efforts to design local, regional, and national responses to the virus are constrained by a lack of information on the extent of the epidemic as well as inaccuracies in newly developed diagnostic tests. In this study we analyze data from testing randomly selected Indiana state residents for infection or previous exposure to SARS-CoV-2 and derive estimates of the statewide COVID-19 prevalence in an attempt to address potential biases arising from nonresponse and diagnostic testing errors.

Author contributions: C.T.Y., P.K.H., and N.M. designed research; C.T.Y. performed research; C.T.Y. analyzed data; C.T.Y. wrote the paper; P.K.H. and N.M. edited the paper; N.M. was principal investigator of the Indiana State contract to perform the study; and P.K.H. and N.M. liaised with the Indiana Department of Health and the Governor's Office.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: cyiannou@iu.edu.

Published January 13, 2021.

as relayed to the dashboard available from The Johns Hopkins University (1).

The study was reviewed by the Indiana University Institutional Review Board. Per the Office of Human Research Protections guidance on coronavirus research (11), this study was deemed a public health surveillance activity exempted from human subjects review. Nevertheless, the IDOH obtained informed consent for this study from each participant.

Statistical Methods

Sample Selection. Selection of study participants was performed randomly via a stratified random-sampling procedure with the 10 IDOH preparedness districts (12) used as strata (Fig. 1). A listing of all Indiana residents was prepared from data obtained from the Indiana State Department of Revenue, for every person who submitted a tax return in the fiscal year 2018 or 2019 plus their dependents. Data were supplemented by information provided by the Indiana Bureau of Motor Vehicles, in cases where information was incomplete or not available. Residents were excluded from sampling if they were less than 12 years of age on the date of the sample selection (22 April 2020), had a non-Indiana address on their tax return, were incarcerated, were deceased, or had no reliable date of birth information.

Endpoints of Interest. The endpoints of interest in the study included having a positive molecular or antibody test. These cases, respectively, address the presence of active disease (molecular test-positive) or previous disease (antibody test-positive). A third endpoint, involving positivity in either or both of these tests, addresses cumulative exposure to the virus (8). In this paper, we address analyses for all three of these endpoints.

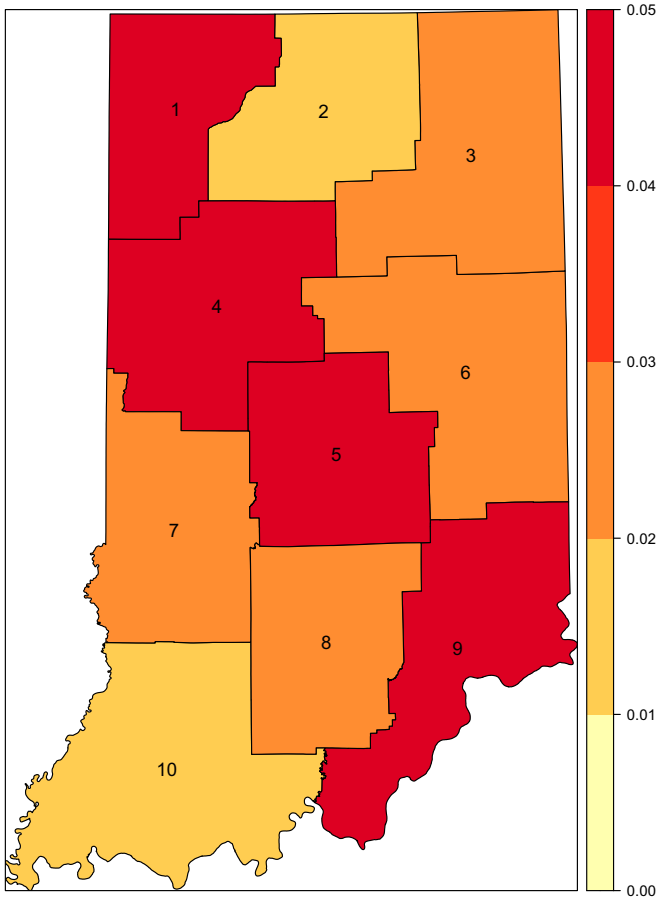

Fig. 1. IDOH preparedness districts in the state of Indiana with heat map corresponding to estimated COVID-19 prevalence (Table 5). Map information from IndianaMap, the largest publicly available collection of Indiana geographic information system map data (9).

Survey Sampling. In survey sampling, inference on the characteristic of interest in the population follows well-established theory (13). The underlying assumption is that the sampling design is "ignorable" (14), which basically means that all factors related to the selection of the samples are accounted for. This is plausible if inclusion of subjects in the study is solely determined by the sampling design. In this case, the prevalence of COVID-19 disease is

$$p = \sum_{i=1}^{I} \frac{N_i p_i}{N} = \sum_{i=1}^{I} w_i p_i,$$

where $w_i = N_i/N$, $i = 1, \ldots, I = 10$ is the fraction of the population in each stratum (IDOH preparedness district) i in the Indiana state population, and $p_i = \sum_{k=1}^{N_i} y_{ik}/N_i$ is the prevalence within each stratum, based on COVID-19–infected ($y_{ik} = 1$) and –uninfected individuals ($y_{ik} = 0$) for $k = 1, \ldots, N_i$. The usual estimate of p is

$$\hat{p} = \sum_{i=1}^{I} \frac{n_i \hat{p}_i}{n} = \sum_{i=1}^{I} \hat{w}_i \hat{p}_i,$$

where n_i and n are, respectively, the number of sampled units within each stratum and the total sample size, while $\hat{p}_i = \sum_{k=1}^{n_i} y_{ik}/n_i$ and $\hat{w}_i = n_i/n$ denote the estimate of the within-stratum prevalence and the stratum fraction in the population, respectively.

Poststratification and Nonresponse. From a survey sampling perspective, sampling weights are determined by the design and can thus be considered deterministic (a priori known and constant). When significant discrepancies exist between the sample fractions n_i/n from the population fractions N_i/N, however, $\hat{p}$ will not be an unbiased estimator of p, as the within-sample estimates will not receive the correct weight. The ignorability assumption is also not plausible when significant nonresponse exists and, in particular, when the likelihood of response is associated with the presence or absence of the characteristic of interest (in our case, exposure to SARS-CoV-2). Adjustments in both cases involve weighting by the inverse probability of selection into the sample (design weights) and poststratification to adjust for known discrepancies between the sample and the population (poststratification weights). In the previous development, we now add a layer $j = 1, \ldots, J$, corresponding to the J-related poststratification groups, and we introduce a nonresponse indicator, R_{ijk}, for individual k in poststratification group j and stratum i, where $R_{ijk} = 1$ if an individual consented for testing and $R_{ijk} = 0$ otherwise. Then, the poststratified estimate of the prevalence is (15),

$$\hat{p}_{ps} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} R_{ijk} w_{ijk} y_{ijk}}{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} R_{ijk} w_{ijk}}, \qquad [1]$$

where y_{ijk} is the SARS CoV-2 infection status for individual k belonging in the poststratification group j in stratum i, and w_{ijk} is the poststratification weight

$$w_{ijk} = \left(\frac{N_i}{n_i}\right)\left(\frac{n_{ij}}{m_{ij}}\right),$$

where n_{ij} is the number of sampled individuals in stratum i and group j, and $m_{ij} = \sum_{k=1}^{n_{ij}} R_{ijk}$ is the number of individuals from that group and stratum actually tested. In this regard, the w_{ijk} provide an adjustment to the inverse probability of sampling (i.e., N_i/n_i) by the poststratification weight n_{ij}/m_{ij}. Poststratification

attempts to correct the sampling weights so that $\hat{p}_{ps}$ gets closer to $\hat{p}$. However, in contrast to sampling weights, poststratification weights are not fixed by design (see, for example, Lu and Gelman) (16), as it is not known at the time of sample selection what the intersection will be between various relevant subgroups in the population and the sample. Consequently, the n_{ij} are random, which in turn means that using them in the process of poststratification is expected to increase the variability of the estimates in contrast to the sampling weights, which are constant. As sampling weights apply equally to all subpopulations, the estimate in [**1**] reduces to

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{I} \left(\frac{N_i}{n_i} \right) \sum_{j=1}^{J} n_{ij} \hat{p}_{ij}, \qquad \textbf{[2]}$$

where $\hat{p}_{ij} = \frac{1}{m_{ij}} \sum_{k=1}^{n_{ij}} R_{ijk} y_{ijk}$ is the estimate of the prevalence in district i and group j. Estimates of COVID-19 prevalence in each district can be produced as

$$\hat{p}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{J} n_{ij} \hat{p}_{ij}, \qquad \textbf{[3]}$$

along with prevalence rates across demographic groups,

$$\hat{p}_{\cdot j} = \frac{\sum_{i=1}^{I} \left(\frac{N_i}{n_i} \right) n_{ij} \hat{p}_{ij}}{\sum_{i=1}^{I} \left(\frac{N_i}{n_i} \right) n_{ij}}. \qquad \textbf{[4]}$$

The intuition of this latter estimate is that the numerator is an estimate of the number of persons infected in demographic group $j = 1, \ldots, J$, while the denominator is an estimate of $\hat{N}_j$, the total number of Indiana residents in group j. See Chen et al. (15) for more details.

When multiple groups are involved, poststratification can be complicated, particularly because of the unknown size of population subgroups resulting from complex interactions of multiple factors. However, in many situations, the marginal counts are known. For example, while we have census data on sex, age, race, and ethnicity in Indiana, we do not know how many non-White, Hispanic individuals live in a certain district in the state. In this case, iterative proportional fitting (17) and raking methods (18) can be used to approximate the probabilities of selection. In this paper, we use iterative proportional fitting (19), where the sample data are weighted so that they match census data in terms of ethnicity (distinguished as study participants of Hispanic versus non-Hispanic ethnicity) and race (categorized as White versus non-White subjects) in the state of Indiana.

Testing Inaccuracy. To address concerns about testing inaccuracy, we performed sensitivity analyses based on various scenarios of testing errors. Given the novelty of the virus, the speed of its worldwide spread and the severity of the resulting pandemic, molecular and antibody tests have been rapidly developed and have generally not been subjected to the usual approval review that diagnostic tests undergo under normal circumstances. Consequently, there are significant concerns about test accuracy. For this reason, both false-positive as well as false-negative errors must be taken into account when calculating final prevalence estimates. Given estimates of false-positive and false-negative rates, the observed prevalence $p^* = P(t = 1)$, the probability of an observed positive test result, based on the true prevalence p follows the equation

$$p^* = p(1 - f_n) + (1 - p) f_p, \qquad \textbf{[5]}$$

where f_p and f_n are, respectively, the estimates of the false-positive and false-negative rates of the test (6, 7). In other words, $f_p\%$ positive test results are added and $f_n\%$ negative tests are subtracted to derive the observed prevalence rate p^*. These methods are applied to estimates of testing accuracy for both the molecular and antibody tests involved in our study, as these enter in the assessment of total disease prevalence. In the IDOH statewide study, depending on the participating laboratory and the timing and location of the sample collection, nasopharyngeal swabs were transferred to the laboratories of Eli Lilly and Company and processed by a laboratory-developed SARS-CoV-2 test (LDT), based on the Centers for Disease Control and Prevention (CDC) primer sets, or to Indiana University Health, where they were processed by the Luminex NxTAG CoV Extended Panel or Roche cobas SARS-CoV-2 test. Blood was transferred to the Mid America Clinical Laboratories for testing using the Abbott IgG test for SARS-CoV-2 IgG Assay. Both the molecular and antibody tests are reported to be highly accurate in limited testing. In our analyses, we assume that the false-positive rate is at most 0.1% for the Luminex assay (20), the Abbott serological test (21), and Lilly's LDT (22) and 2% for the Roche cobas test (23, 24). We also assume that the false-negative rate is 3% for the Luminex assay (20), 0.3% for the Roche cobas test (24), and 0.4% for the Abbot assay given research available by the companies and independent laboratories (21, 25). We assign false-negative rates on the antibody test on the lower end of the reported ranges, because high false-negative rates are primarily observed early in the infection (25), when presumably positive cases would be detected by the PCR test (resulting in lower overall false-positive rates). The expression in [**5**] is used unchanged in the analyses of the antibody test results. To account for the two different RT-PCR tests, we modify the expression in [**5**] slightly as

$$p^* = \sum_{t=1}^{2} \delta_t \left[p(1 - f_{nt}) + (1 - p) f_{pt} \right], \qquad \textbf{[6]}$$

where δ_t, $t = 1, 2$, is an indicator of the RT-PCR test involved in each test, while f_{nt} and f_{pt} are, respectively, the false-negative and false-positive rates associated with each of the two molecular tests. In the analyses of cumulative disease prevalence, we simply add the two expressions, where now the prevalence associated with antibody testing is related to the excess prevalence of previous SARS-CoV-2 exposure, among people without active disease (see prior elicitation in *Bayesian Analysis* for more details).

Bayesian Analysis. To bring all components of the analysis together and properly propagate the error through them, we use Bayesian methods. See, for example, Qian et al. (6), Chen et al. (15), and Gelman and Carpenter (7) for related ideas. The model is

$$y_{ij} \sim \text{Binomial}(m_{ij}, p_{ij}),$$

where $y_{ij} = \sum_{k=1}^{n_{ij}} R_{ijk} y_{ijk}$ reflects m_{ij} test results in stratum i and group j. To account for multiple RT-PCR and antibody tests, the above model is modified as $y_{ij} \sim \text{Binomial}(m_{ij}, p^*_{ij})$, where p^*_{ij} is defined in [**5**] or [**6**] as appropriate, in order to account for the different tests as described in the model above. Prevalence of cumulative disease exposure is estimated as the sum of current disease and the excess of cases with previous exposure to SARS CoV-2 but without active disease. In this case, prevalence of prior exposure is determined as the difference of cumulative disease and prevalence of active disease (with the constraint that it be greater or equal to zero). We impose beta priors on the true prevalence p_{ij} and the false-negative and false-positive rates of each test, i.e.,

$$p_{ij} \sim \text{Beta}(a, b)$$

$$f_n \sim \text{Beta}(a_n, b_n)$$

$$f_p \sim \text{Beta}(a_p, b_p).$$

Table 1. Confirmed COVID-19 cases during the 2-wk period of 15 to 29 April 2020 and total cases since the start of the pandemic in Indiana along with populations of the 10 IDOH preparedness districts in the state, used in the elicitation of priors for seroprevalence and rates of active disease

IDOH district	Cases in 15 to 29 April	Total cases	District population
1	1,293	2,394	811,393
2	619	1,049	657,419
3	591	883	741,028
4	1,341	1,478	379,126
5	3,429	8,292	1,866,050
6	521	1,092	626,343
7	102	264	277,283
8	329	744	382,115
9	463	1,259	463,370
10	187	380	487,755
Total	8,875	17,835	6,691,882

The hyperparameters in the beta priors are obtained as the numbers of false-positive and false-negative tests performed in the various laboratory studies consulted in the prior elicitation (see references in *Testing Accuracy*). For an aggregated list of such references also see the Foundation of Innovative New Diagnostics (26).

For the prior distribution of the COVID-19 prevalence, we consider information available in the IDOH dashboard on the number of daily confirmed cases. We then determine the parameters in the prior distribution so that the 95% CI of the beta distribution covers 1 to 55 times the number of cases reported by IDOH, divided by each district's population (a crude measure of disease prevalence). This is loosely based on results from the Stanford study (4). To calculate the resulting prior distributions on cumulative and active COVID-19 disease prevalence, we work backward to ensure that the average prevalence remains at the levels reported by IDOH (listed in Table 1). To account for differences in prevalence rates among ethnic and racial subgroups, we follow the CDC estimates, which assign triple the risk for infection with SARS-CoV-2 to individuals with Hispanic ethnicity and, on average, double the risk among non-White racial groups, compared to White persons of non-Hispanic ethnicity (27). For example, there have been 2,394 confirmed COVID-19 cases in District 1 (Table 1), out of a population of 811,393. Assuming a 3:3:2:1 ratio of cases among Hispanic non-Whites and Whites and non-Hispanic non-Whites and Whites (27), we estimate that non-Hispanic Whites have 0.2% observed prevalence, while Hispanic Whites have 0.6% prevalence in the district in order for the combined prevalence to be 0.3% (the ratio of total cases 2,394 over the 811,393 residents in the districts). The beta parameters for the prevalence prior among non-Hispanic Whites is $a = 1.34$ and $b = 37.09$, resulting in prior 95% CI for the prevalence, between 0.20 and 11.06% (or 55 times the observed prevalence). The beta parameters corresponding to Hispanic Whites (and Hispanic non-Whites) are $a = 1.28$ and $b = 10.40$, corresponding to prevalence bounds equal to three times those among non-Hispanic Whites. We also consider active disease, based on the number of confirmed cases reported in the 2 wk between 15 and 29 April 2020, the day the test was completed in this study (Table 1). In the case of District 1 again, where 1,293 cases were reported in that period, this process results in a beta prior distribution with parameters $a = 1.32$ and $b = 21.77$, for White residents of Hispanic origin, resulting in a 95% prior CI for the prevalence between 0.33 and 17.9%.

Postsampling Simulations. To properly assess the variability of sample sizes in the four demographic subgroups (i.e., Hispanic/non-Hispanic, White/non-White), we have performed simulations, generating repeated sequences of the n_{ij} sampled observations from a multinomial distribution with probabilities N_{ij}/N_i and total sample size $n_i = \sum_{j=1}^{J} n_{ij}$, where the population sizes N_{ij} were obtained from the iterative proportional fitting procedure discussed in *Poststratification and Nonresponse*. We carried out 1,000 such simulations in each of the three separate analyses described in *Results*. The simulated n_{ij} counts were used in the calculations involved in Eqs. **2**–**4** above.

All analyses were performed within the R environment (28). Bayesian inference was carried out using the package RStan (29). Iterative proportional fitting was implemented through the package mipfp (19). Data management was performed with the package dplyr (30), and maps were generated through the packages maps (31) and sp (32). Survey estimates were produced with the package survey (33). All code and data summaries used in these analyses are posted on GitHub (https://github.com/cyiannou/IDOH-STUDY).

Results

Characteristics of the Sample. The selection of Indiana residents was performed according to a stratified random sample based on the 10 IDOH preparedness districts (12) (Fig. 1). There had been about 11,000 confirmed cases reported by IDOH by 20 April 2020 (1, 10) for a crude prevalence estimate of 0.16% in a state of about 6.7 million people (34) (Table 1). A sample of 5,000 residents was calculated to provide an estimate of the prevalence that would a have margin of error of less than 1%,

Table 2. Description of the sampling design

District	District population	Sample population	Exclusions				Total exclusions	Number sampled	Number of tests	Number positive
			No DOB	< 12 y	Deceased	Incarcerated				
1	763,039	2,397	60	436	59	3	558	1,839	380	10
2	644,674	2,025	38	407	47	4	496	1,529	287	0
3	723,066	2,272	39	414	44	5	502	1,770	448	7
4	338,900	1,065	22	181	23	0	226	839	209	2
5	1,834,537	5,763	94	1,102	102	12	1,310	4,453	1,170	37
6	574,535	1,805	15	264	47	2	328	1,477	327	9
7	247,406	778	14	133	14	0	161	617	156	4
8	328,984	1,034	17	165	20	2	204	830	212	1
9	447,988	1,408	33	247	28	0	308	1,100	204	9
10	464,186	1,459	23	272	27	3	325	1,134	232	6
Total	6,367,315	20,006	355	3,621	411	31	4,418	15,588	3,625	85

District population numbers in the state of Indiana were supplied by state agencies for the purposes of sampling and do not correspond to population estimates provided by the US Census Bureau. DOB, date of birth.

Table 3. Relevant sample characteristics

	Sample data (n = 3,625)		Census data (n = 6,691,878)	
Factor	Frequency	Percentage	Frequency	Percentage
Race				
Non-White	279	7.70	878,567	13.13
White	3,346	92.30	5,813,315	86.87
Ethnicity				
Hispanic/Latino	80	2.21	474,572	7.09
Not Hispanic	3,545	97.79	6,217,310	92.91

even under the extreme scenario of a 15% prevalence, the upper limit considered following estimates of unreported cases in the Stanford study (4). To account for nonresponse, a sample of 20,006 residents was selected with probability proportional to the population size of each Indiana IDOH district. Out of these, 4,418 sampled individuals were excluded because they were younger than 12 y on 22 April 2020, were incarcerated, were deceased, or did not have reliable date of birth data, resulting in a final sample of 15,588 residents. From these, 3,658 were tested and 3,625 had at least one available molecular or antibody test present in the database (Table 2). Demographic information on the 3,625 tested subjects are shown in Table 3.

Prevalence Estimates. Unadjusted and adjusted (poststratified) prevalence estimates are shown in Table 4. Point estimates plus Wald and exact binomial CIs are presented with respect to the unadjusted estimates. The statistics presented in the case of adjusted (poststratified) estimates were derived from the empirical distribution resulting from 1,000 simulations (medians of the empirical distribution generated via simulation along with 95% empirical CIs of the 2.5th and 97.5th quantile of the empirical distribution).

Unadjusted Prevalence Estimates. Using the data in Table 2, the point estimate of the total disease prevalence is $\hat{p} = 2.33\%$, with a 95% Wald-type CI of 1.83 to 2.82%. An exact binomial 95% CI is 1.86 to 2.87% (Table 4). These unadjusted estimates involved use of the sampling weights but no other adjustment.

Adjusted Prevalence Estimates. Representation of important demographic subgroups in the sample, such as race and ethnicity, can be contrasted with census data in the state (2). From consideration of these data (Table 3), it appears that people of Hispanic ethnicity and non-White Indiana residents are underrepresented in the sample. Given that ethnicity and race have been identified to be related with COVID-19 prevalence (35), we performed poststratification adjustments as described earlier in *Statistical Methods*. The results, which also account for inaccuracies in the testing, are shown in Table 4 for antibody tests, RT-PCR tests, and the combined measure of cumulative exposure to COVID-19 disease. Adjusting for underrepresentation by non-White individuals and people of Hispanic ethnicity, as well as imperfect testing, resulted in a revised estimate of seroprevalence of 2.60% (95% empirical CI: 2.08 to 3.35%) based on the antibody test, a prevalence of active COVID-19 disease of 1.81% (1.46 to 2.25%) based on the RT-PCR test, and an estimate of cumulative disease exposure of 3.58% (3.03 to 4.18%). In all cases, the adjusted estimates were higher than the unadjusted estimates (Table 4). Results from the analysis of cumulative COVID-19 prevalence are shown in Fig. 2. From Fig. 2, it is evident that the adjustment involving poststratification and consideration of possible erroneous diagnostic tests has resulted in a substantial right shift of the center of the prevalence distribution to the right of the usual stratified (unadjusted) estimate of the prevalence.

District-Level Estimates. Prevalence estimates within each district are shown in Table 5. A heat map, corresponding to the cumulative prevalence rates (rightmost column in Table 5), is shown in Fig. 1. The highest cumulative disease exposure rates were seen in District 5 (the district that includes Indianapolis and surrounding counties), District 1 (located at the northwest corner of the state close to Chicago, IL), and District 9 (the district bordering the tristate area between Indiana, Ohio, and Kentucky and the large metropolitan areas of Cincinnati and Louisville). Very high rates were also observed in District 4, home to a concentrated epidemic surrounding meat-packing plants in the area. Point estimates and 95% empirical CIs are presented in Table 5.

Estimates by Demographic Group. We also produce estimates by demographic groups, broken down into dichotomous classifications of individuals as Hispanic or non-Hispanic and White versus non-White. These are presented in Table 6. The conclusion from these data is that individuals of Hispanic ethnicity and, secondarily, non-White persons, are disproportionately affected by the epidemic, as has been reported previously (36, 37).

Discussion

Statewide testing for infection with the SARS-CoV-2 virus was performed in Indiana between 25 and 29 April 2020. This study was the first randomized statewide testing study for infection with SARS-CoV-2 undertaken in the United States. In the paper describing the results of this study, Menachemi et al. reported an overall disease prevalence estimate of 2.79% (8). Their estimates were adjusted for ethnicity, race, and age, but no allowance was made for possible inaccuracies in the estimates resulting from erroneous diagnostic test results. The revised estimate of 3.58%, resulting after taking into consideration both poststratification for demographic factors as well as possible testing errors, suggests that the cumulative disease prevalence reported by Menachemi et al. may have been somewhat underestimated.

The adjustment of prevalence estimates through poststratification is important because racial and ethnic groups are disparately affected by the COVID-19 epidemic (36, 37) but are frequently underrepresented in surveys (38). In this reanalysis of the Indiana statewide sampling data, poststratification adjustment for Hispanic ethnicity and non-White race and corrections for imperfect testing resulted in substantial increases over the unadjusted estimates (Table 4). The reason for this revision is likely the much higher prevalence rates seen among non-Whites and individuals of Hispanic ethnicity in the sample (8) and possible false-negative test results in both the molecular and antibody tests, even for the relatively small error levels considered.

Table 4. Prevalence estimates

	Seroprevalence			RT-PCR test positivity			Cumulative exposure		
Type of estimate	Point estimate	95% CI		Point estimate	95% CI		Point estimate	95% CI	
Unadjusted (Wald)	1.52	1.11	1.92	1.30	0.92	1.67	2.33	1.83	2.82
Unadjusted (exact)	1.52	1.14	1.98	1.30	0.95	1.72	2.33	1.86	2.87
Poststratified (empirical)	2.60	2.08	3.35	1.81	1.46	2.25	3.58	3.03	4.18

STATISTICS

Table 5. Prevalence estimates adjusted for poststratification and test inaccuracies by IDOH preparedness district

	Seroprevalence			RT-PCR test positivity			Cumulative exposure		
District	Point estimate	95% CI		Point estimate	95% CI		Point estimate	95% CI	
1	3.68	1.98	6.39	2.31	1.11	4.11	4.42	2.69	7.08
2	1.29	0.51	2.86	0.84	0.29	1.91	1.83	0.98	3.30
3	1.25	0.58	2.44	1.31	1.32	0.65	2.32	1.29	3.33
4	2.27	0.80	5.35	2.75	1.10	5.76	4.37	2.24	7.61
5	3.89	2.69	5.92	2.11	1.38	3.11	5.00	3.76	6.54
6	1.33	0.59	2.79	1.98	1.06	3.28	2.83	1.74	4.47
7	0.89	0.33	2.29	1.31	0.54	2.63	2.15	0.99	4.29
8	1.37	0.56	3.07	0.82	0.33	1.80	2.06	1.14	3.66
9	4.04	2.04	7.28	1.48	0.57	3.12	4.09	2.32	6.81
10	0.94	0.37	2.12	1.52	0.71	2.78	1.95	1.06	3.63

With the revised estimate of COVID-19 disease prevalence of 3.58%, the total number of cases in Indiana as of the end of April is estimated to be 241,044 in a state of about 6.7 million people (34). Menachemi et al. (8) estimated the number of COVID-19 cases to be 187,802 individuals, corresponding to a prevalence estimate of 2.79% of the Indiana population. Excluding the 85 positive cases identified in this study, there were 17,756 confirmed COVID-19 cases in the state from the start of the epidemic up to 29 April 2020 (10). The estimate by Menachemi et al. suggests an almost 11-fold difference between confirmed and estimated number of COVID-19 cases in Indiana as of 29 April 2020, the date the testing was completed. Adjusting for possible test inaccuracies, our revised estimate of the fold difference between confirmed and total COVID-19 cases is 13.6 (95% empirical CI: 11.5 to 15.8). These numbers represent a much lower order of magnitude than the 55-fold difference (95% CI: 26 to 95) reported in the Stanford study (4).

We also produced district-specific estimates, subjected to both poststratification adjustments and analyses to account for imperfect testing. The district-level cumulative disease-exposure estimates are highly variable, ranging from just under 2% in districts 2 and 10, to 5% in District 5. The highest levels of COVID-19 disease were observed mostly in localities including or being proximal to large urban areas like Marion County (Indianapolis) and surrounding counties (District 5); the counties close to Chicago, IL (District 1); and the cities of Louisville and Cincinnati, OH (District 9) (Fig. 1). On the other hand, the highest rate of active disease was seen in District 4, reflecting an emerging concentrated epidemic in meat-packing plants in the district. Having the ability to track the evolution of the epidemic and to differentiate between regions within a state is helpful for mobilizing state resources efficiently to areas of most acute need.

As reported elsewhere (27, 36), the burden of COVID-19 is not equal across demographic groups. In our study, non-White Indiana residents and, in particular, persons of Hispanic ethnicity, had by far the highest rates of disease prevalence. The differences are particularly stark when compared with prevalence rates among White non-Hispanic individuals.

A major advantage of Bayesian modeling, in addition to providing a unified platform for carrying out the entire analysis, is the ability to incorporate all available evidence in the model. In this manner, important nuances, such as previous versus current disease can be detected, with significant subepidemics observed among minority populations and urban centers as early as the end of April, less than 2 mo after the first case was identified in Indiana. The analysis also detected an emerging epidemic in a district where a superspreader event occurred in meat-packing plants in late April. This is a great strength of this approach.

At the same time, our study has a number of limitations. The most important of these involves potential bias resulting from the low response rate. We attempted to adjust for underrepresentation in the sample among important demographic groups by poststratifying our data using census information. It should be acknowledged, however, that these adjustments make the implicit assumption that missing data are missing at random (MAR) (39). To clarify, under MAR, a tested Hispanic White person would have the same chance of having a positive test as a member of this subgroup who did not respond to the invitation for testing. However, if nonresponders have different prevalence of COVID-19 disease than responders with similar characteristics, even poststratified analyses will result in estimates with unknown bias. A positive bias in this study is improbable, however. While it is possible that persons who were motivated to be tested, and might have had higher disease prevalence, responded preferentially to the invitation, almost 40% of those testing positive in the study reported having no symptoms (8). Consequently, there is weak evidence that symptomatic individuals tried to avail themselves of a free testing opportunity during a time of testing scarcity. By contrast, nonresponders, such as Hispanic and non-White residents, would be expected to have higher rather

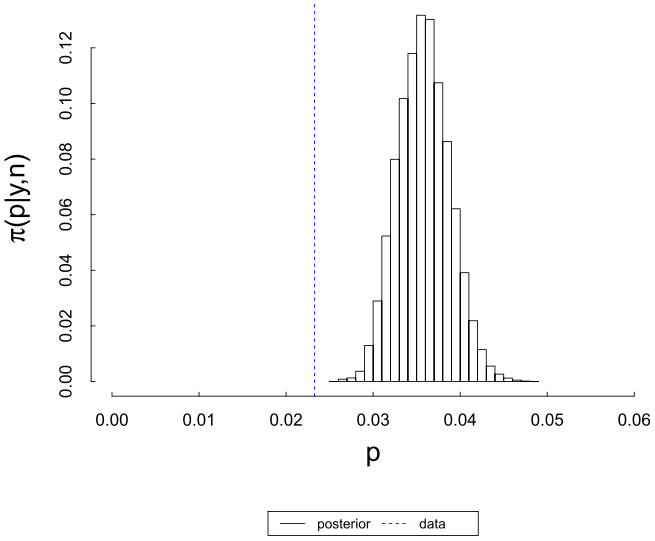

Fig. 2. Cumulative exposure to SARS-CoV-2 in Indiana at the end of April 2020. The histogram shows the posterior distribution of the cumulative prevalence adjusted for nonresponse and imperfect testing. The dashed line shows the usual stratified estimate of the prevalence without any further adjustments.

Table 6. Medians and 95% CIs of the posterior distribution for overall disease prevalence in the four demographic groups in the state of Indiana

	Seroprevalence			RT-PCR test positivity			Cumulative exposure		
Demographic	Point estimate	95% CI		Point estimate	95% CI		Point estimate	95% CI	
Hispanic/non-White	9.15	3.58	20.49	5.54	2.41	11.60	10.80	5.99	19.50
Hispanic/White	9.75	5.08	17.59	6.56	3.60	10.90	12.00	7.39	18.30
Non-Hispanic/non-White	5.08	3.11	8.52	3.22	1.92	5.25	6.95	4.76	9.69
Non-Hispanic/White	1.57	1.19	2.09	1.16	0.89	1.50	2.30	1.89	2.78

than lower disease prevalence (8, 40), so we are unlikely to have overestimated the prevalence of the disease. On the other hand, it is hard to imagine much higher numbers of cases in the state, say in the order of the Stanford or USC studies. If, for example, the true number of cases in Indiana were 55 times the number of confirmed cases reported by the state, the total number of COVID-19 cases in Indiana by 29 April 2020 would have been 1,031,800 (95% CI: 487,760 to 1,782,200), a prevalence of 15.3% (95% CI: 7.2 to 26.5%). A cumulative prevalence of up to one in four Indiana residents is virtually impossible given data reported by state agencies. For example, there had been 6,445 COVID-19–related intensive care unit (ICU) and non-ICU hospital admissions reported in the state from 6 March 2020, the date of the first confirmed case of COVID-19 in Indiana, and 29 April 2020, when testing for the present study was completed (41). If the true number of cases in the state were over 1 million, even hospitalization rates similar to severe influenza would be expected to result in much higher numbers of hospitalizations during the same period. Simple calculations using CDC estimates of hospitalizations during the 2017 to 2018 severe flu season (42) (dividing the number hospitalized by the total number of influenza cases and multiplying the result by the point estimate of total cases in Indiana per the Stanford study) result in about 18,611 expected hospitalizations (95% CI: 8,798 to 32,146) in Indiana during the same period. This is almost three to five times higher than the number reported by state agencies. Thus, the supposition of a much lower or greater caseload than estimated in these analyses is not aligning with study or state data.

Another limitation of these data is a potential bias resulting from even small levels of error in the diagnostic testing. If the test specificity (true-negative) rate is not virtually 100%, then there could be enough false-positive tests to put in doubt a significant portion of the already small number of observed positive test results. For example, even a 1% false-positive rate would suggest that, on average, 36 of the positive tests observed in our study could be due to false-positive results, erasing almost three-quarters of the observed positive RT-PCR or antibody test results. If the false-positive rate were even higher, it would render the study virtually uninterpretable. Fortunately, false-positive error rates do not seem to be of great concern in the tests used in this study, based on manufacturer and independent laboratory evaluations (20–22, 24, 25). Of less concern is a lower sensitivity (higher false-negative) rate scenario. While crucial clinically, false-negative errors are not expected to materially affect our prevalence estimates as they correspond to latent COVID-19–positive individuals who, in the early stages of the epidemic, are expected to be few in numbers. For example, if the false-negative rate of a test were 1% and the overall disease prevalence were 3%, then on average, only 1 to 2 cases out of about 3,600 would be missed as false-negative results. The combined effect of poststratification adjustments and corrections for false-positive and false-negative rates in the tests used in the study was a revision in the estimate of the statewide prevalence of infection with SARS-CoV-2 from 2.33 to 3.58% (more than 50% higher than the stratified—unadjusted—estimate and almost 30% higher than 2.79%, the estimate reported by Menachemi et al.) (8).

A final limitation and potential source of bias is the construction of the original population from which the sample was drawn. As described in *Statistical Methods*, the basis for constructing the sampling frame were data from the Department of Revenue, including residents who had submitted tax returns in the 2018 and 2019 tax years, supplemented by Indiana Bureau of Motor Vehicle data. This population excluded anyone who had not submitted a tax return in the past 2 y and did not have a current driver's license. If these individuals had higher- or lower-than-average disease prevalence, their exclusion from sampling would result in underestimation or overestimation of the final statewide estimate.

These concerns, however, do not detract from the utility and significance of our study. The alternative to random sampling, even with all of the possible biases and caveats listed here, can result in seriously questionable estimates. In their paper, Menachemi et al. (8) also report testing among 898 persons through outreach in the African American and Hispanic communities in Indianapolis. In this nonrandom sample, 22.8% of those tested had a positive PCR test, and an additional 5.8% had a positive antibody test without testing positive on the PCR test (8), resulting in a cumulative disease prevalence of just under 30%, an estimate that is well outside any credible levels for the entire state, or the region surrounding Indianapolis, for that period.

We conclude that our analysis, of the first randomized survey sampling and testing of infection with SARS-Cov-2, despite a number of potential sources of error and uncertainty in the estimates, is useful as a guide when calculating the prevalence of a relatively rare disease in this population and time period.

Data Availability Anonymized data have been deposited in GitHub (https://github.com/cyiannou/IDOH-STUDY/).

ACKNOWLEDGMENTS. We thank the thousands of Indiana residents who participated in the study, without whom this unique and important effort would have been impossible. We acknowledge the support of the Indiana Governor's Office in marshalling the resources needed for successfully conducting this study, along with the IDOH and numerous state agencies and private organizations, who contributed meaningfully to this endeavor. We thank the Indiana Management Performance Hub in particular, which was integral in generating the population of Indiana residents from which the sample was drawn. We are privileged to work alongside these professionals.

1. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
2. X. He *et al.*, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
3. N. Sood *et al.*, Seroprevalence of SARS-CoV-2–specific antibodies among adults in Los Angeles county, California on April 10-11, 2020. *J. Am. Med. Assoc.* **323**, 2425–2427 (2020).
4. E. Bendavid *et al.*, COVID-19 antibody seroprevalence in Santa Clara County, California. medRxiv [Preprint] (2020). http://doi.org/10.1101/2020.04.14.20062463 (Accessed 8 January 2021).
5. US Food and Drug Administration, Coronavirus disease 2019 (COVID-19) emergency use authorizations for medical devices. https://www.fda.gov/medical-devices/emergency-situations-medical-devices/emergency-use-authorizations. Accessed 2 July 2020.

6. S. S. Qian, J. M. Refsnider, J. A. Moore, G. R. Kramera, All tests are imperfect: Accounting for false positives and false negatives using Bayesian statistics. *Heliyon* **6**, e03614 (2020).
7. A. Gelman, B. Carpenter, Bayesian analysis of tests with unknown specificity and sensitivity. *J. R. Stat. Soc. Ser. C Appl. Stat.* **69**, 1269–1283 (2020).
8. N. Menachemi *et al.*, Population prevalence of COVID-19 from an Indiana statewide random sample. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 960–964 (2020).
9. Indiana Geographic Information Council, Indianamap. https://www.indianamap.org/. Accessed 6 November 2020.
10. Indiana State Department of Health, State of Indiana COVID-19 dashboard. https://www.coronavirus.in.gov/2393.htm. Accessed 25 June 2020.
11. Office of Human Research Protections, OHRP guidance on coronavirus. https://www.hhs.gov/ohrp/regulations-and-policy/guidance/ohrp-guidance-on-covid-19/index.html. Accessed 7 November 2020.
12. Indiana State Department of Health, Indiana public health preparedness districts. https://www.in.gov/isdh/17944.htm. Accessed 8 June 2020.
13. W. G. Cochran, *Sampling Techniques* (Wiley, New York, NY, ed. 3, 1977).
14. D. B. Rubin, Inference and missing data (with discussion). *Biometrika* **63**, 581–592 (1976).
15. C. X. Chen, T. Lumley, J. Wakefield, The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spat. Spatiotemporal Epidemiol.* **11**, 33–43 (2014).
16. H. Lu, A. Gelman, A method for estimating design-based sampling variances for surveys with weighting, post-stratification and raking. *J. Off. Stat.* **19**, 133–151 (2003).
17. W. E. Deming, F. F. Stephan, On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* **11**, 427–444 (1940).
18. J. Deville, C. Sarndal, O. Sautory, Generalizing raking procedures in survey sampling. *J. Am. Stat. Assoc.* **88**, 1013–1020 (1993).
19. J. Barthélemy, T. Suesse. mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. *J. Stat. Softw.*, **86**, 1–20, 2018.
20. J. H.-K. Chen *et al.*, Clinical performance of the Luminex NxTAG CoV extended panel for SARS-CoV-2 detection in nasopharyngeal specimens of COVID-19 patients in Hong Kong. *J. Clin. Virol.*, 10.1128/JCM.00936-20 (2020).
21. A. Bryan *et al.*, Performance characteristics of the Abbott Architect SARS-CoV-2 IgG assay and seroprevalence in Boise, Idaho. *J. Clin. Microbiol.* **58**, e00941–20 (2020).
22. A. K. Nalla *et al.*, Comparative performance of SARS-CoV-2 detection assays using seven different primer-probe sets and one assay kit. *J. Clin. Microbiol.* **58**, e00557 (2020).
23. M. Poljak *et al.*, Clinical evaluation of the cobas SARS-CoV-2 test and a diagnostic platform 1 switch during 48 hours in the midst of the CoVid-19 pandemic. *J Clin Microbiol.* **58**, e00599–20 (2020).
24. J. A. Lieberman1 *et al.*, Comparison of commercially available and laboratory developed assays for in vitro detection of SARS-CoV-2 in clinical laboratories. *J. Clin. Microbiol.* **58**, e00821 (2020).
25. M. S. Tang *et al.*, Clinical performance of two SARS-CoV-2 serologic assays. *Clin. Chem.* **66**, 1055–1062 (2020).
26. The Foundation for Innovative New Diagnostics, SARS-CoV-2 diagnostics: Performance data. https://www.finddx.org/covid-19/dx-data/. Accessed 21 September 2020.
27. Centers for Disease Control and Prevention, COVID-19 hospitalization and death by race/ethnicity. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html. Accessed 21 September 2020.
28. R Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2020).
29. Stan Development Team, RStan: the R interface to Stan (R Package Version 4.0.2, Stan Development Team, 2020).
30. H. Wickham, R. François, L. Henry, K. Müller, dplyr: A Grammar of data manipulation (R Package Version 1.0.2, 2020). https://cran.r-project.org/package=dplyr. Accessed 8 January 2021.
31. R. A. Becker, A. R. Wilks, R. Brownrigg, T. P. Minka, A. Deckmyn, maps: Draw geographical maps (R Package Version 3.3.0, 2018). https://cran.r-project.org/package=maps. Accessed 8 January 2021.
32. R. S. Bivand, E. Pebesma, V. Gomez-Rubio, *Applied Spatial Data Analysis with R* (Springer, New York, NY, ed. 2, 2013).
33. T. Lumley, *Complex Surveys A Guide to Analysis Using R* (Wiley, 2010).
34. United States Census Bureau, QuickFacts Indiana: Population estimates. https://www.census.gov/quickfacts/IN. Accessed 1 July 2019.
35. E. Williamson *et al.*, OpenSAFELY: Factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients. medRxiv [Preprint] (2020). http://doi.org/10.1101/2020.05.06.20092999 (Accessed 8 January 2021).
36. A. Van Dorn, R. E. Cooney, M. L. Sabin, COVID-19 exacerbating inequalities in the US. *Lancet* **395**, 1243–1244 (2020).
37. J. A. W. Gold *et al.*, Characteristics and clinical outcomes of adult patients hospitalized with COVID-19 - Georgia, March 2020. *MMWR Morb. Mortal. Wkly. Rep*. **69**, 545–550 (2020).
38. K. L. Schneider, M. A. Clark, W. Rakowski, K. L. Lapane, Evaluating the impact of nonresponse bias in the behavioral risk factor surveillance system (BRFSS). *J. Epidemol. Community Health* **66**, 290–295, 2012.
39. R. J. A. Little, D. B. Rubin, *Statistical Analysis with Missing Data* (Wiley, ed. 2, 2002).
40. G. A. Millett *et al.*, Assessing differential impacts of COVID-19 on black communities. *Ann. Epidemiol.* **47**, 37–44 (2020).
41. Regenstrief Institute, Regenstrief COVID-19 dashboard. https://www.regenstrief.org/covid-dashboard/. Accessed 21 September 2020.
42. Centers of Disease Control and Prevention National Center for Immunization and Respiratory Diseases (NCIRD), Estimated influenza illnesses, medical visits, hospitalizations, and deaths in the United States – 2017–2018 influenza season. https://www.cdc.gov/flu/about/burden/2017-2018.htm. Accessed 22 November 2020.