A primer on Bayesian estimation of prevalence of COVID-19 patient outcomes
Xiang Gao and Qunfeng Dong

DEPARTMENT OF MEDICINE, STRITCH SCHOOL OF MEDICINE, LOYOLA UNIVERSITY CHICAGO, 2160 S. FIRST AVENUE, MAYWOOD, IL 60153, USA

Corresponding Author: Qunfeng Dong, PhD, Director, Center for Biomedical Informatics, Loyola University Chicago, Maywood, IL 60153, USA (qdong@luc.edu)

Received 9 October 2020; Editorial Decision 22 October 2020; Accepted 29 October 2020

ABSTRACT
A common research task in COVID-19 studies often involves the prevalence estimation of certain medical outcomes. Although point estimates with confidence intervals are typically obtained, a better approach is to estimate the entire posterior probability distribution of the prevalence, which can be easily accomplished with a standard Bayesian approach using binomial likelihood and its conjugate beta prior distribution. Using two recently published COVID-19 data sets, we performed Bayesian analysis to estimate the prevalence of infection fatality in Iceland and asymptomatic children in the United States.

Key words: COVID-19, SARS-CoV-2, Bayesian, conjugate prior, infection fatality risk, asymptomatic

INTRODUCTION
Many COVID-19 studies are interested in estimating the prevalence of certain medical outcomes of interest. Typically, the prevalence was reported as a point estimate accompanied by a 95% confidence interval (95% CI). For example, in a study recently published by Gudbjartsson et al., the authors estimated the prevalence of COVID-19 deaths in Iceland, obtaining the infection fatality risks of 0.1% (95% CI 0.0–0.3%), 2.4% (95% CI 0.6–6.2%), and 11.2% (95% CI 3.6–24.0%) for those 70 years old or younger, those between 70 and 80 years of age, and those older than 80, respectively. In another recent study published by Sola et al., the authors estimated the prevalence of infected children without any COVID-19 symptoms for multiple regions in the United States, showing a pooled asymptomatic prevalence of 0.65% (95% CI 0.47–0.83%).

There are three main limitations with the traditional biostatistical methods used to obtain the above estimations. First, the above studies only obtained point estimates for the prevalence inferred from the available data. Although point estimates may be the most likely values of the unknown prevalence, values other than the point estimates may also have a non-negligible high probability. Since there always exists uncertainty associated with any inferred values for prevalence, the uncertainty should be ideally measured by a probability distribution that assigns a precise probability to every possible value of the unknown prevalence (ie, values with higher
Table 1. Bayesian analysis of two published COVID-19 data sets

| Study | Age groups (years old) | Death (y) | Infection (N) | Prior Beta(a, b) | Posterior Beta(a + y, b + N − y) | Posterior median (95% credible interval), % |
|-------|------------------------|-----------|---------------|-----------------|------------------|------------------------------------------|
| Infection fatality rates in Iceland | 0–70        | 3         | 3012          | Beta(1)         | Beta(4, 3010)    | 0.12 (0.04–0.29)                        |
|       | 70–80                  | 3         | 128           | Beta(1)         | Beta(4, 126)     | 2.84 (0.85–6.65)                       |
|       | >80                    | 4         | 38            | Beta(1)         | Beta(5, 35)      | 11.87 (4.30–24.22)                     |

| Study | Regions | ASX (y) | Infection (N) | Prior Beta(a, b) | Posterior Beta(a + y, b + N − y) | Posterior median (95% credible interval), % |
|-------|---------|---------|---------------|-----------------|------------------|------------------------------------------|
| Asymptomatic (ASX) children in U.S. | West    | 120     | 15311         | Beta(1)         | Beta(121, 15192) | 0.79 (0.66–0.94)                       |
|       | Midwest | 40      | 5217          | Beta(1)         | Beta(41, 5178)   | 0.78 (0.56–1.04)                       |
|       | South   | 49      | 8354          | Beta(1)         | Beta(50, 8306)   | 0.59 (0.44–0.78)                       |
|       | Northeast | 41     | 4159          | Beta(1)         | Beta(42, 4119)   | 1.00 (0.73–1.33)                       |

APPLICATION TO COVID-19 DATA

We have applied the above binomial and beta model to perform Bayesian analysis on two recently published COVID-19 data sets (Table 1). Since we did not have any prior knowledge on the infection fatality rate or the asymptomatic prevalence, we used a noninformative beta prior (ie, both its shape parameters, $a$ and $b$, were set to the value of 1). We then plugged in the necessary numbers to calculate the posterior distributions by updating the parameters of the beta distributions (Table 1). For example, for the age group 0–70 years old in Iceland, there were three deaths ($y$) out of a total of 3012 infections ($N$), so the posterior probability distribution of the infection fatality risk for this age group is $beta(3 + 1, 1 + 3012 − 3)$. Similarly, out of a total of 15,311 infected children ($N$) in the West region of United States, 120 were asymptomatic ($y$), so the posterior distribution for the prevalence of asymptomatic children in the West region of U.S. is $beta(1 + 120, 1 + 15,311 − 120)$.

After obtaining the posterior distributions (ie, the beta distributions with updated parameters), we can visualize the distributions by randomly sampling from them and plotting the samples. Figures 1 and 2 depict the posterior distributions for infection fatality rates in Iceland and the prevalence of asymptomatic children in the United States, respectively, which provide a complete probabilistic landscape for those parameters. Besides plotting, the posterior distributions are also often characterized by summary statistics, for
Figure 1. The posterior probability densities of infection fatality rate for different age groups in Iceland: (A) 0–70, (B) 70–80, and (C) >80.

Figure 2. The posterior probability density of the prevalence of asymptomatic children in four different US regions: (A) West, (B) Midwest, (C) South, and (D) Northeast.
example, medians and 95% credible intervals (Table 1). It is important to note that contrary to confidence intervals, credible intervals represent the likely ranges of the true values of the unknown parameter.6 We provided an example R11 programming script (Supplementary File S1) for plotting the posterior distributions and calculating the summary statistics. Although our current estimations were based on noninformative prior probability distributions for prevalence, informative priors can be used if relevant information is available. In fact, our current estimates can become informative priors for future updates using the same Bayesian framework.

DISCUSSION
Bayesian analyses are often perceived as complicated. It is true that applying Bayesian analyses may require highly customized modeling procedures. For example, we have recently published COVID-19 related studies using Bayesian approaches,12,13 which required (1) developing customized likelihood functions and (2) the estimation of the posterior distributions by MCMC. However, as illustrated above via the reanalysis of the two published COVID-19 data sets, estimating prevalence can be easily achieved using a simple Bayesian model based on binomial likelihood and its beta conjugate prior, which is mathematically straightforward and well applicable for prevalence estimation in real-world data analysis. As researchers around the world are gathering more and more COVID-19 data for estimating the prevalence of various medical outcomes, we hope that Bayesian approaches will be widely utilized. In our own experience, the presented Bayesian model is a stepping stone for beginners to appreciate the power of Bayesian approaches before learning more complicated models (eg, Bayesian hierarchical modeling) and computational techniques (eg, MCMC).

SUPPLEMENTARY MATERIAL
Supplementary material is available at Journal of the American Medical Informatics Association online.

AUTHORS’ CONTRIBUTIONS
X.G. performed data analysis. Q.D. drafted the manuscript. Both conceived the project.

CONFLICT OF INTEREST STATEMENT
None declared.

REFERENCES
1. Gudbjartsson DF, Norddahl GJ, Melsted P, et al. Humoral immune response to SARS-CoV-2 in Iceland. N Engl J Med 2020; 383 (18): 1724–34.
2. Sola AM, David AP, Roske KW, Baba A, Ramirez-Avila L, Chan DK. Prevalence of SARS-CoV-2 infection in children without symptoms of coronavirus disease 2019. JAMA Pediatr 2020; doi: 10.1001/jamapediatrics.2020.4095.
3. Morey RD, Hoekstra R, Rouder JN, et al. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev 2016; 23 (1): 103–23.
4. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016; 31 (4): 337–50.
5. Lunn AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Educ Anaesth Crit Care Pain 2008; 8 (6): 221–3.
6. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. 3rd ed. London: Chapman & Hall, 2013.
7. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 2000; 10 (4): 325–37.
8. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS Book: A Practical Introduction to Bayesian Analysis. Boca Raton, FL: Chapman and Hall/CRC; 2013.
9. Plummer M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: 3rd International Workshop on Distributed Statistical Computing (DSC 2003). Vienna, Austria, 2003, 124.
10. Carpenter B, Gelman A, Hoffman MD, et al. Stan: a probabilistic programming language. J Stat Soft 2017; 76 (1). doi: 10.18637/jss.v076.i01
11. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
12. Gao X, Dong Q. A Bayesian framework for estimating the risk ratio of hospitalization for people with comorbidity infected by the SARS-CoV-2 virus. JAMIA 2020; https://doi.org/10.1093/jamia/ocaa246.
13. Dong Q, Gao X. Bayesian estimation of the seroprevalence of antibodies to SARS-CoV-2. JAMIA Open 2020. https://doi.org/10.1093/jamiaopen/ocaa049.