The Annals of Applied Statistics
2022, Vol. 16, No. 1, 436–459
https://doi.org/10.1214/21-AOAS1499

BAYESIAN ADJUSTMENT FOR PREFERENTIAL TESTING IN ESTIMATING INFECTION FATALITY RATES, AS MOTIVATED BY THE COVID-19 PANDEMIC

BY HARLAN CAMPBELL[1,a], PERRY DE VALPINE[2], LAUREN MAXWELL[3], VALENTIJN M. T. DE JONG[4], THOMAS P. A. DEBRAY[4,5], THOMAS JAENISCH[3,6] AND PAUL GUSTAFSON[1,b]

[1]*Department of Statistics, University of British Columbia,* [a]*harlan.campbell@stat.ubc.ca,* [b]*gustaf@stat.ubc.ca*
[2]*Department of Environmental Science, Policy, and Management, University of California*
[3]*Heidelberg Institute for Global Health, Heidelberg University Hospital*
[4]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University*
[5]*Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University*
[6]*Department of Epidemiology, Colorado School of Public Health*

A key challenge in estimating the infection fatality rate (IFR), along with its relation with various factors of interest, is determining the total number of cases. The total number of cases is not known not only because not everyone is tested but also, more importantly, because tested individuals are not representative of the population at large. We refer to the phenomenon whereby infected individuals are more likely to be tested than noninfected individuals as "preferential testing." An open question is whether or not it is possible to reliably estimate the IFR without any specific knowledge about the degree to which the data are biased by preferential testing. In this paper we take a partial identifiability approach, formulating clearly where deliberate prior assumptions can be made and presenting a Bayesian model which pools information from different samples. When the model is fit to European data obtained from seroprevalence studies and national official COVID-19 statistics, we estimate the overall COVID-19 IFR for Europe to be 0.53%, 95% C.I. = [0.38%, 0.70%].

1. Introduction. If someone is infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the pathogen that causes COVID-19, how likely is that person to die of COVID-19? This simple question is surprisingly difficult to answer.

The "case fatality rate" (CFR) is a common measure that quantifies the mortality risk in a certain population and is given by the ratio of deaths (D) over confirmed cases (CC) during a specific time period. However, because many COVID-19 cases are never diagnosed, the CFR almost certainly overestimates the true lethality of the virus. Instead, the better answer is captured by the infection fatality rate (IFR) (Kobayashi et al. (2020), Wong et al. (2013)). The IFR, also a simple ratio, differentiates itself from the CFR by considering all cases, including the asymptomatic, undetected and misdiagnosed infections in the denominator. For instance, if 20 individuals die of the disease in a population with 1000 infections, then the IFR is $20/1000 = 0.02 = 2\%$.

Evidently, a key challenge in calculating the IFR is determining the true total number of cases. The total number of cases (C) is not known because not everyone is tested in the population (P). A naïve estimate of the IFR might take this into account by simply considering the number of tests (T) and estimating the number of cases as: $\mathrm{C} \approx (\mathrm{CC}/\mathrm{T}) \times \mathrm{P}$. However, diagnostic tests are often selectively initiated such that tested individuals are not representative of the population at large.

Received October 2020; revised May 2021.
Key words and phrases. Selection bias, partial identification, evidence synthesis.

In most countries/jurisdictions, those with classic COVID-19 symptoms (e.g., fever, dry cough, loss of smell or taste) are much more likely to be tested than those without symptoms. Due to this "severity bias," the reported number of cases likely includes mostly people whose symptoms were severe enough to be tested and excludes the vast majority of those who are mildly impacted or asymptomatic. Even when testing is made equally available to all individuals, there is potential for bias if people who have reason to believe they are infected are more likely to volunteer to be tested (e.g., Bendavid et al. (2020)). We refer to the phenomenon whereby infected individuals are more likely to be tested than noninfected individuals as "preferential testing." (Hauser et al. (2020) and others use the term "preferential ascertainment.")

If the degree of preferential testing in a particular sample is of known magnitude, bias adjustment can be achieved by appropriately altering the estimated rate of infection and its uncertainty interval. However, the degree of preferential testing is typically unknown and likely highly variable across different jurisdictions. An open question is whether or not it is possible to reliably estimate the IFR without any specific information about the degree to which the data are biased by preferential testing (Q1). And, if we have some samples for which testing is representative and others which are subject to some unknown bias from preferential testing, is it better to use only the representative data or to combine both kinds of data in a joint analysis (Q2)? In this paper we address these two important questions by considering a Bayesian hierarchical model for estimation of the IFR. We demonstrate with an application of the model to European data from seroprevalence studies and national official COVID-19 statistics.

Bayesian models have been previously used in similar situations. For example, Presanis et al. (2009) use Bayesian inference to estimate the severity of pandemic H1N1 influenza. More recently, Rinaldi and Paradisi (2020) and Hauser et al. (2020) use Bayesian models for disease dynamics in order to estimate the severity of COVID-19. To address the issue of preferential testing bias, Hauser et al. (2020) apply susceptible-exposed-infected-removed (SEIR) compartmental models to age-stratified data and, in order to establish parameter identifiability, assume that all cases of infected patients aged 80 years and older are confirmed cases. The Bayesian model we propose is more general and allows one to obtain appropriate point and interval estimates for the IFR with varying degrees of prior knowledge about the magnitude of preferential testing and the distribution of other explanatory factors (e.g., age, healthcare capacity).

This paper is structured as follows. In Section 2 we introduce required notation, discuss distributional assumptions and review key issues of identifiability. In Section 3 we formulate our Bayesian model, and in Section 4 we describe how the model can be scaled for larger populations and can incorporate covariates. In Section 5 we present a simulation study, and in Section 6 we present an analysis of COVID-19 data from Europe. We conclude in Section 7 with a return to the primary questions of interest (Q1 and Q2).

2. Notation, distributions, and issues of (un)identifiability.

2.1. *Notation and distributions*. Suppose we have data from K independent groups (i.e., countries or jurisdictions) from a certain fixed period of time. For group $k = 1, \ldots, K$, let:

- P_k be the population size (i.e., the total number of individuals at risk of infection);
- T_k be the total number of people tested;
- CC_k be the total number of confirmed cases resulting from the tests, and
- D_k be the total number of observed deaths attributed to infection.

We do not observe the following latent variables. For the kth group, let:

- C_k be the total number of infected people (cases) in the population;
- IR_k be the true infection rate (proportion of the population which is infected), which is the expected value of C_k/P_k, and
- IFR_k be the true underlying infection fatality rate (IFR) which is the expected value of D_k/C_k.

Therefore, we assume that

$$C_k \sim \text{Binom}(P_k, IR_k), \quad \text{and} \tag{1}$$

$$D_k|C_k \sim \text{Binom}(C_k, IFR_k), \tag{2}$$

where in the kth group the unknown number of infections, C_k, and the known number of deaths, D_k, each follow a binomial distribution. Note that there are latent variables on both the left-hand side and the right-hand side of (1).

For each group, CC_k is recorded, instead of C_k. Even in the absence of preferential testing, CC_k will be smaller than C_k because not everyone is tested. The goal is to draw inference on the relationship between the number of deaths, D, and the number of cases, C, having only data on D, CC, P and T. This is particularly challenging since the number of confirmed cases in each group may be subject to an unknown degree of preferential testing.

In the absence of any preferential testing, if one assumes that the population sizes are finite, then the number of confirmed cases will follow a hypergeometric distribution (Prochaska and Theodore (2018)). The hypergeometric distribution describes the probability of CC_k confirmed cases amongst T_k tests (without any individuals being tested more than once) from a finite population of size P_k that contains exactly C_k cases. Wallenius' *noncentral* hypergeometric is a generalization of the hypergeometric distribution whereby testing is potentially biased with either cases or noncases more likely to be tested (Fog (2008)). We, therefore, consider the distribution of $CC_k|C_k$ as following a noncentral hypergeometric (NCHG) distribution:

$$CC_k|C_k \sim \text{NCHG}(C_k, P_k - C_k, T_k, \phi_k), \tag{3}$$

where the degree of preferential testing corresponds to the ϕ_k noncentrality parameter (see Appendix (Section A.1) for details about the NCHG distribution). When $\phi_k > 1$, cases (i.e., infected individuals) are more likely to be tested than noncases (i.e., noninfected individuals); when $\phi_k < 1$, cases are less likely to be tested than noncases. When $\phi_k = 1$, we have that the probability of being tested is equal for both cases and noncases, and the NCHG distribution reduces to the standard hypergeometric distribution. In this parameterization the ϕ_k parameter can be interpreted as an odds ratio, that is, the odds of a case being tested vs. the odds of a noncase being tested.

The distribution of the confirmed cases depends on the actual infection rate (C/P) and the testing rate (T/P) but does not depend on the infection fatality rate (D/C). In other words, we assume that the conditional distribution of (CC|C, T, P, D) is identical to the conditional distribution of (CC|C, T, P). This assumption is similar to the assumption of "nondifferential" exposure misclassification in measurement error models and may or may not be realistic; see Smedt et al. (2018). If across the K different groups those groups with higher ϕ_k values also tend to have higher IFR_k values, then one will inevitably obtain biased estimates because the IFR_k and ϕ_k are considered a priori independent. The same logic applies to the IR_k and IFR_k which are also a priori independent.

Also, note that the members of set D_k are not a subset of the members of set CC_k. While D_k is a subset of C_k, and CC_k is a subset of $C_k \cap T_k$, D_k is not necessarily a subset of CC_k. For example, in the seroprevalence study data for Luxembourg, which we consider in Section 6 (see Supplementary Material—Table 1, row 3, Campbell et al. (2022)), we have $CC_k = 23$

confirmed cases out of $T_k = 1214$ tests. There are $D_k = 93$ deaths out of a population of $P_k = 615{,}729$. Evidently, D_k is not a subset of CC_k. Furthermore, the assumption that $\phi_k = 1$ for this Luxembourg data implies that the 1214 tested individuals were not any more or less likely to be infected than those in the general population. However, note that there is no requirement that the tested individuals have the same risk of death as those in the general population. To be clear, no distributional assumptions will be violated if, within the kth group, individuals with a higher probability of death (e.g., the elderly) are more likely to be tested than those with a lower probability of death (e.g., young, healthy individuals).

2.2. *Partial identifiability.* Given the assumptions detailed above, for each of the K groups there are three unknown parameters (latent states), IR_k, IFR_k and ϕ_k that must be estimated for every two observed quantities (D_k/P_k and CC_k/T_k). This indicates that a unique solution will not be attainable without additional external data or prior information.

The problem at hand is sufficiently rich and complex that forming intuition about the information-content of the data is challenging. In the Appendix (Section A.2) we consider, in depth, an asymptotic argument for *partial identifiability*. We determine that, depending on the range and heterogeneity in the degree of preferential testing across groups, the data can contribute substantial information about the IFR. Data from any single group may only be weakly informative about the IFR, in the sense that only lower and upper bounds for the IFR are estimable. However, we show that, in some circumstances, there is very considerable sharpening of information when these bounds are combined across groups, provided it is a priori plausible that the IFR heterogeneity across groups is modest.

3. A Bayesian model for small-P data. We describe a Bayesian model, which assumes standard Gaussian random-effects, allowing both the infection rate (IR) and infection fatality rate (IFR) to vary between groups. Bayesian models are known to work well for dealing with partially identifiable models; see Gustafson (2010). Consider the following random-effects model:

$$\mathrm{g}(\mathrm{IFR}_k) \sim \mathcal{N}(\theta, \tau^2), \quad \text{and} \tag{4}$$

$$\mathrm{g}(\mathrm{IR}_k) \sim \mathcal{N}(\beta, \sigma^2), \tag{5}$$

for $k = 1, \ldots, K$, where θ is the parameter of primary interest, τ^2 represents between group IFR heterogeneity, β represents the mean g(infection rate), σ^2 describes the variability in infection rates across the K groups and g() is a given link function. Note that, alternatively, a simpler fixed-effects version of the model arises by setting $\tau = 0$ such that $\mathrm{g}(\mathrm{IFR}_k) = \theta$, for $k = 1, \ldots, K$.

We will adopt the complimentary log-log link function (cloglog) for g(), though there are other sensible choices, including the logit and probit functions. Our choice of the cloglog function facilitated the creation of parameter-transformed samplers for efficient sampling (see Section 1 in Supplementary Material (Campbell et al. (2022))).

Putting together the assumptions for $p(D_k|\mathrm{IFR}_k, C_k)$, $p(CC_k|T_k, P_k, C_k, \phi_k)$ and $p(C_k|P_k, \mathrm{IR}_k)$, defined in Section 2.1 along with prior distributions, Bayes' Law takes the form

$$\begin{aligned} p\big((\theta, \tau^2, \beta, \sigma^2, \mathrm{C}, \mathrm{IFR}, \mathrm{IR}, \phi)|\mathrm{data}\big) &\propto p\big(\mathrm{data}|\theta, \tau^2, \beta, \sigma^2, \mathrm{C}, \mathrm{IFR}, \mathrm{IR}, \phi\big) \\ &\quad \times p\big(\theta, \tau^2, \beta, \sigma^2, \mathrm{C}, \mathrm{IFR}, \mathrm{IR}, \phi\big) \end{aligned} \tag{6}$$

$$= \left(\prod_{k=1}^{K} p(\mathrm{D}_k|\mathrm{IFR}_k, \mathrm{C}_k) p(\mathrm{CC}_k|\mathrm{T}_k, \mathrm{P}_k, \mathrm{C}_k, \phi_k) p(\mathrm{C}_k|\mathrm{P}_k, \mathrm{IR}_k) \right.$$
$$\left. \times\, p\big(\mathrm{IFR}_k|\theta, \tau^2\big) p\big(\mathrm{IR}_k|\beta, \sigma^2\big) \right) \times p(\theta) p\big(\tau^2\big) p(\beta) p\big(\sigma^2\big) p(\phi).$$

We are left to define prior distributions for the unknown parameters: θ, τ^2, β, σ^2 and ϕ. Our strategy for priors on IR and IFR is to not only assume uninformative priors for the mean of IFR and of IR and for the variance of IR but also a strongly informative prior favouring small values for the variance of IFR. This strategy reflects the assumption that the infection fatality rate varies across jurisdictions much less than the infection rate itself (especially after accounting for population level sources of heterogeneity; see Section 4.2). Uniform and half-Normal priors are set accordingly: $\mathrm{g}^{-1}(\theta) \sim \mathrm{Uniform}(0, 1)$; $\mathrm{g}^{-1}(\beta) \sim \mathrm{Uniform}(0, 1)$; $\sigma \sim \text{half-}\mathcal{N}(0, 1)$ and $\tau \sim \text{half-}\mathcal{N}(0, \eta^2)$, where $\eta = 0.1$.

The only remaining component is $p(\phi)$. Our strategy for a prior on the degree of preferential testing is to assume that cases are more likely to be tested than noncases (i.e., $\phi_k > 1$), that all values of ϕ_k are equally likely across jurisdictions and that there is an upper bound, $1 + \gamma$, on the degree of preferentiality. For the upper bound parameter, γ, we assume an exponential prior such that

$$\phi_k|\gamma \sim \mathrm{Uniform}(1, 1 + \gamma), \quad \text{for } k = 1, \ldots, K; \quad \text{and} \quad \gamma \sim \mathrm{Exp}(\lambda).$$

We, therefore, assume that the uniform range of possible values for ϕ_k is itself unknown. This hierarchy allows one to specify a very "weakly informative" prior for the degree of preferential testing. For instance, setting $\lambda = 0.05$ implies that, a priori, a reasonable value for the ϕ_k odds ratio is about 6 (infected individuals are about six times more likely to be tested than those uninfected) and could range anywhere from about 3 to 14. (When $\lambda = 0.05$, the median of the unconditional distribution for ϕ_k is 6.4 with a wide interquartile range of 2.7 to 14.3.)

In some scenarios we might have some groups for which ϕ_k is known and equal to 1 (i.e., have data from some samples where testing is known to be truly random). Without loss of generality, suppose this subset is the first k' studies such that, for $k = 1, \ldots, k'$, we have $\phi_k = 1$. We will use this approach in the European data analysis (Section 6) in which we assume ϕ_k is known and equal to 1 for data from representative seroprevalence studies.

We must emphasize that the performance of any Bayesian estimator will depend on the choice of priors and that this choice can substantially influence the posterior when few data are available (Berger (2013), Lambert et al. (2005)). The priors described here represent a scenario where there is little to no a priori knowledge about the θ, β and ϕ model parameters. Inference would no doubt be improved should more informative priors be specified based on probable values for each of these parameters. We will consider the impact of priors in the simulation study in Section 5, where we look to different values for λ and η.

We must also emphasize that, due to the partial identifiability issues (Section 2.2), a delicate trade-off may exist between the priors for the τ and ϕ parameters. For instance, if large values of τ are made a priori plausible (i.e., if η is large), then the posterior estimates of the ϕ parameters may be driven downward toward 1 (due to the $\gamma \sim \mathrm{Exp}(\lambda)$ prior). A relatively homogeneous across-group IFR can be central to identifiability and, as such, the aforementioned "fixed-effects" version of the model (essentially equivalent to fixing $\tau = 0$) may be more feasible in situations when identification is particularly challenging (e.g., when $k' = 0$ and/or when there is very little prior knowledge about the θ, β and ϕ model parameters). On the other hand, in situations when identification is less of a concern (e.g., when k' is relatively large relative to K and/or when there is substantial and reliable prior information),

setting a priori limitations on τ may be detrimental if the true heterogeneity in infection fatality rates across groups is high and meaningful. In the meta-analysis literature it is well known that adding studies within an analysis that are too heterogeneous, without accounting for the heterogeneity, can actually result in greater uncertainty rather than greater precision (Sutton et al. (2007)).

4. A Bayesian model for large-P data.

4.1. *Distributional approximations*. When populations are sufficiently large, two simplifications to the model are desirable. First, we will replace the NCHG distribution with a binomial distribution as follows:

$$\mathrm{CC}_k \sim \mathrm{Binom}\big(\mathrm{T}_k, 1-(1-\mathrm{C}_k/\mathrm{P}_k)^{\phi_k}\big) \tag{7}$$

for $k = 1, \ldots, K$. This simplification[1] alleviates the need for writing custom samplers for the NCHG distribution for certain MCMC software (e.g., Stan, nimble) and also provides additional familiarity to researchers who may not be accustomed to working with the NCHG distribution. Second, we can dispense with the need to sample the C_k latent variables by replacing the above distribution for CC_k with

$$\mathrm{CC}_k \sim \mathrm{Binom}\big(\mathrm{T}_k, 1-(1-\mathrm{IR}_k)^{\phi_k}\big). \tag{8}$$

For any sufficiently large P_k, this simplification will make little to no difference. Then, since the distributions of C_k and $\mathrm{D}_k|\mathrm{C}_k$ are both binomials (see (1) and (2)), we have that, unconditionally,

$$\mathrm{D}_k \sim \mathrm{Binom}(\mathrm{P}_k, \mathrm{IFR}_k \times \mathrm{IR}_k). \tag{9}$$

Note that, in (7) and (8) above, the ϕ_k parameter no longer corresponds to an odds ratio, yet the interpretation is similar. Starting from (8), the odds ratio (OR) describing the association between testing status and infection status is

$$\log(\mathrm{OR}) = \log\big(1-(1-\mathrm{IR})^{\phi}\big) - \phi \times \log(1-\mathrm{IR}) - \log(\mathrm{IR}) + \log(1-\mathrm{IR}).$$

For fixed IR, approximating this with a Taylor series in $\log(\phi)$, about zero, gives: $\log(\mathrm{OR}) \approx c_{\mathrm{IR}} \log(\phi)$, where $c_{\mathrm{IR}} = -\log(1-\mathrm{IR})/\mathrm{IR}$. Note that $c_{\mathrm{IR}} \to 1$ as $\mathrm{IR} \to 0$. Therefore, in the rare-infection realm, ϕ is indeed approximately the odds ratio for testing and infection status.

4.2. *Including group-level covariates*. The proposed model can be expanded to include factors of interest specified as covariates at the group level, resembling what is commonly done in a meta-regression analysis (Thompson and Higgins (2002)). Covariates included for analysis might be metrics that are correlated with the probability of infection, with the probability of being tested, with the accuracy of the test and/or with the probability of dying from infection.

For instance, suppose that $X_{[1]k}, \ldots, X_{[h]k}$ are h different group-level covariates that explain the kth group's infection rate and that $Z_{[1]k}, \ldots, Z_{[q]k}$ are q different covariates that explain the kth group's IFR. Then these can be incorporated as follows:

$$\mathrm{g}(\mathrm{IR}_k) \sim \mathcal{N}\big(\beta + \beta_1 X_{[1]k} + \ldots + \beta_h X_{[h]k}, \sigma^2\big), \tag{10}$$

$$\mathrm{g}(\mathrm{IFR}_k) \sim \mathcal{N}\big(\theta + \theta_1 Z_{[1]k} + \ldots + \theta_q Z_{[q]k}, \tau^2\big). \tag{11}$$

[1]Recall that a hypergeometric distribution is asymptotically equivalent to a binomial distribution, and, while this particular binomial parameterization does not emerge from the limit of the NCHG distribution, it is a reasonable approximation. We could have alternatively substituted the NCHG distribution with the known Gaussian asymptotic approximation to the NCHG (Stevens (1951)). However, the Gaussian approximation requires solving quadratic equations and, therefore, might not actually make things simpler; see Sahai and Khurshid (1995).

Age is a key factor for explaining the probability of COVID-19-related death (O'Driscoll et al. (2020)). One might, therefore, consider median age of each group as a predictor for the IFR or perform analyses that are stratified by different age groups (Onder, Rezza and Brusaferro (2020)). The latter strategy has, for instance, been recommended to make accurate predictions for respiratory infections (Pellis et al. (2020)). With regards to the infection rate, time since first reported infection or time between first reported infection and the imposition of social distancing measures might be predictive (Anderson et al. (2020)).

4.3. *MCMC.* For the large-P model, Markov chain Monte Carlo (MCMC) mixing can be slow because different combinations of ϕ_k, cloglog(IR_k) and cloglog(IFR_k) can yield similar model probabilities. This is related to the identifiability issues discussed in the Appendix (Section A.2). Standard Gibbs sampling (e.g., as implemented with JAGS (Kruschke (2014))) will be inefficient in many situations. To improve mixing and reduce computational time, we wrote the model in the nimble package (de Valpine et al. (2017)) which supports an extension of the modeling language used in JAGS and makes it easy to configure samplers and provide new samplers. Details of the MCMC implementation for nimble are presented in the Supplementary Material (Campbell et al. (2022)). We also implemented the large-P model in the popular Stan package which employs Hamiltonian MCMC algorithms (Carpenter et al. (2017)).

5. Simulation study.

5.1. *Design.* We conducted a simulation study in order to better understand the operating characteristics of the proposed model. Specifically, we wished to evaluate the frequentist coverage of the credible interval for θ and to investigate the impact of choosing different priors.

As emphasized in Gustafson and Greenland (2009), the average frequentist coverage of a Bayesian credible interval, taken with respect to the prior distribution over the parameter space, will equal the nominal coverage. This mathematical property is unaffected by the lack of identification. However, the variability of coverage across the parameter space is difficult to anticipate and could be highly affected by the choice of prior. For example, we might expect that, in the absence of preferential testing (i.e., when $\gamma = 0$), coverage will be lower than the nominal rate. However, if this is the case, coverage will need to be higher than the nominal rate when $\gamma > 0$ so that the "average" coverage (taken with respect to the prior distribution over the parameter space) is nominal overall.

We simulated datasets with $K = 20$ and $k' = 8$. For $k = 1, \ldots, 8$, population sizes were obtained from a NegBin(20,000, 1) distribution with a mean and standard deviation of 20,000 and for $k = 9, \ldots, 20$; population sizes were obtained from a NegBin(200,000, 1) distribution (with mean and standard deviation equal to 200,000). Parameter values were as follows: $\theta = \text{cloglog}(0.02) = -3.90$, $\beta = \text{cloglog}(0.20) = -1.50$, $\tau^2 = 0.005$ and $\sigma^2 = 0.25$. The testing rate for each population was obtained from a Uniform(0.01, 0.10) distribution so that the proportion of tested individuals in each population ranged from 1% to 10%. We considered eight values of interest for γ: 0, 0.5, 1, 2, 4, 12, 32, 64 (for simulation), and three different values of interest for both λ and η: 0.05, 0.1 and 0.5 (for estimation). The number of confirmed cases (CC_k) were simulated from Wallenius' NCHG distribution, as detailed in Section 2.1. The 12 "unknown" ϕ_k values, for $k = 9, \ldots, 20$, were simulated from a Uniform$(1, \gamma + 1)$ distribution. Note that, with high γ levels, the vast majority of tests will be positive (when $\gamma = 32$, positivity is about 72%; when $\gamma = 64$, positivity is about 81%).

We fit three models to each unique dataset: $M1$, $M2$ and $M3$. All three models follow the same large-P framework detailed in Section 4.1, but each considers a different subset of the data:

- The $M1$ model uses only data from the groups for which ϕ_k is unknown, that is, $\{P_k, T_k, CC_k \text{ and } D_k\}$ for $k = 9, \ldots, 20$ ($k' = 0$, and $K = 12$);
- The $M2$ model considers the data from all 20 groups, that is, $\{P_k, T_k, CC_k, \text{ and } D_k\}$ for $k = 1, \ldots, 20$ ($k' = 8$, and $K = 20$), and
- The $M3$ model uses only data from the groups for which ϕ_k is known and equal to 1, that is, $\{P_k, T_k, CC_k, \text{ and } D_k\}$ for $k = 1, \ldots, 8$ ($k' = 8$, and $K = 8$).

To be clear, the $M2$ and $M3$ models make the assumption of (correctly) known $\phi_k = 1$ for $k = 1, \ldots, 8$.

We simulated 1100 unique datasets (i.e., 1100 unique sets of values for $\{P_k, T_k, D_k, CC_{k,\gamma}\}$, for $k = 1, \ldots, K$, and $\gamma = \{0, 0.5, 1, 2, 4, 12, 32, 64\}$) and, for each dataset, fit the three different models. (See Table 2 in the Appendix for an example of a "single" unique dataset.) We specifically chose to conduct 1100 simulation runs so as to keep computing time within a reasonable limit while also reducing the amount of Monte Carlo standard error (MCSE) to a reasonably small amount. For looking at coverage with $1-\alpha = 0.90$, MCSE will be approximately $\sqrt{0.90(1-0.90)/1100} < 0.01$; see Morris, White and Crowther (2019).

For each unique dataset the $M1$ and $M2$ models were fit 72 times ($= 3 \times 3 \times 8$): with λ assuming one of the three values of interest, with η assuming one of the three values of interest and with one of the eight different sets of CC_k numbers (for $k = 9, \ldots, 20$) corresponding to the nine γ values of interest. The $M3$ model was fit three times for each unique dataset, with η assuming one of the three values of interest. For each model fit we recorded the posterior median estimate of $\text{icloglog}(\theta)$, the width of the 90% highest posterior density (HPD) CI for θ and whether or not the 90% HPD CI contained the target value of $\text{cloglog}(0.02) = -3.90$.

For each simulation scenario we used Stan to obtain a minimum of $N_{\text{MC}} = 18,000$ MCMC draws from the posterior (a total from three independent chains, with 20% burn-in and thinning of 5). We recorded the Gelman–Rubin test statistic, $\hat{R}$ (Gelman, Rubin et al. (1992), Brooks and Gelman (1998)), and if this statistic was $\hat{R} > 1.05$, the MCMC sampling was discarded and was restarted anew with twice the number of MCMC draws, up to a maximum of $N_{\text{MC}} = 288,000$. If, even after four restarts, with $N_{\text{MC}} = 288,000$, we obtained $\hat{R} > 1.05$, a convergence/mixing failure was recorded and the result was simply discarded.

5.2. *Results*. Figure 1 plots the simulation study results. The $M3$ model, which only considers data from those groups where testing is known to be representative/random, appears to obtain a average point estimate for $\text{icloglog}(\theta)$ of approximately 0.02, as desired for all three values of η. In contrast, the $M1$ model, which only considers data from those groups where the degree of preferential testing is unknown, obtains average point estimates for $\text{icloglog}(\theta)$ far above and far below the target value of 0.02, depending on η, λ and γ. (Note that many results for $M1$ are so large/small that they are outside the limits of the plot.) The $M2$ model, which makes use of all the data, obtains point estimates of approximately 0.02 for all positive values of γ, when λ is sufficiently small (i.e., $\lambda \leq 0.1$) for all three values of η. When $\lambda = 0.5$, the $M2$ model tends to underestimate $\text{icloglog}(\theta)$ when η and/or γ are large.

Coverage for models $M2$ and $M3$ appears to be highly dependent on η. With $\eta = 0.1$, the $M2$ and $M3$ models obtain coverage of approximately 90%, as desired for all $\gamma > 0$ values considered. With $\eta = 0.05$, coverage is ever so slightly less than the desired 90% level, and, when $\eta = 0.5$, coverage is higher than the desired 90% level. The results from the $M1$ model show that, for small values of λ and η (i.e., for $\lambda \leq 0.5$ and $\eta < 0.5$), coverage is at or above 90% for the entire range of γ values. This suggests that appropriate coverage may be achievable, even when $k' = 0$ and when in the presence of a substantial and unknown amount of preferential testing.

The credible interval width results from the $M2$ and $M3$ models indicates that, for a wide range of γ values, the $M2$ model (which makes use of all the data) is preferable to the $M3$

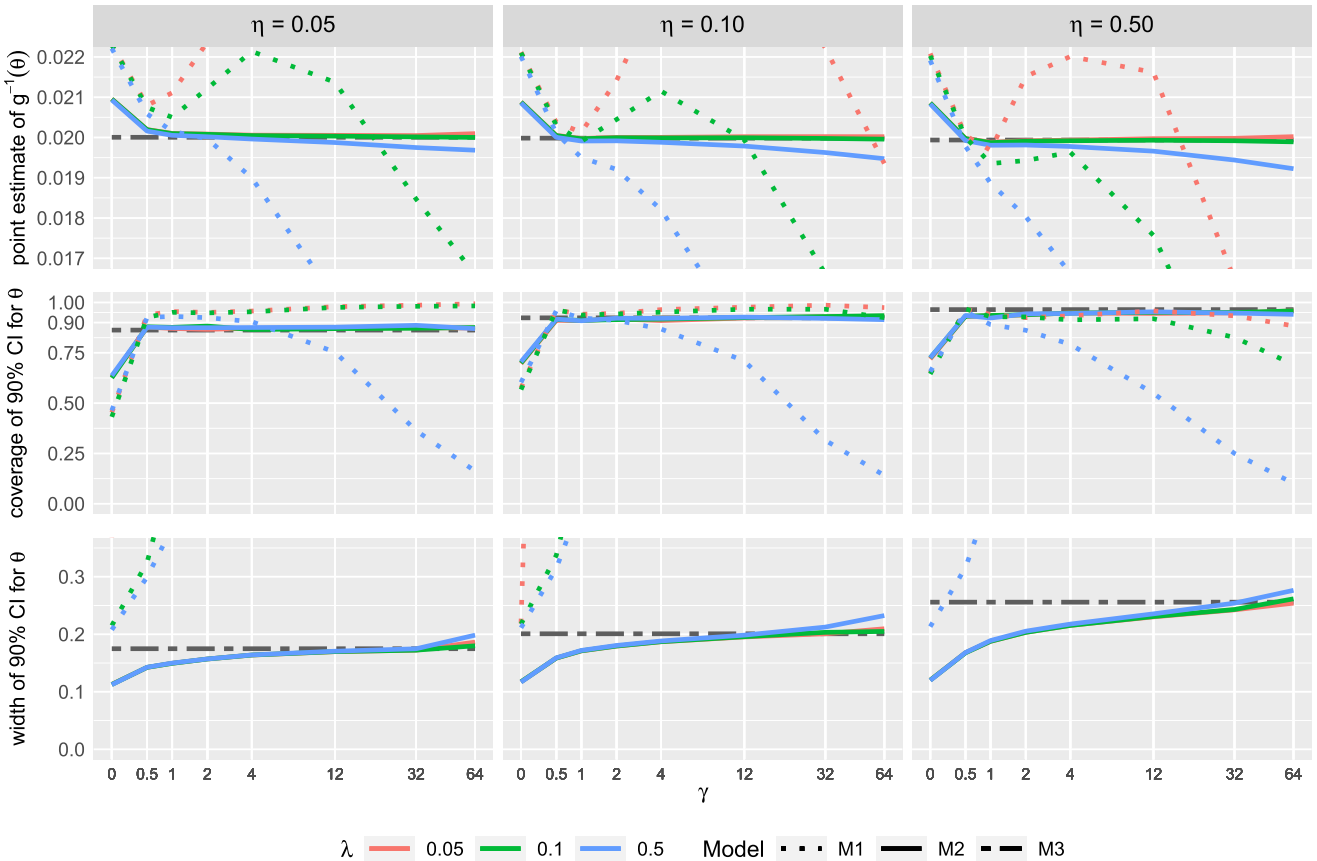

FIG. 1. *Results from the simulation study. The top row plots average point estimate obtained for* $\text{icloglog}(\theta)$*, the middle row plots the coverage of the* 90% *HPD CI and the bottom row plots the average width of the* 90% *HPD CI. Each column of panels corresponds to a different level of* η*. To be clear, each different value for* γ *corresponds to a different upper bound on the degree of preferential testing in the simulated data. Different values for* λ *and* η *correspond to different prior specifications.*

model (which uses data only from those groups where testing is known to be representative/random). However, there is a limit to the "added value" that the "nonrepresentative" data provide. For example, for $\gamma > 12$ and $\eta = 0.1$, $M3$ intervals are narrower, compared to $M2$ intervals (for all values of λ).

Overall, the interval width is much much narrower for $M2$ relative to $M1$. This confirms that the $k' = 8$ representative samples are very valuable for reducing the uncertainty around θ. (Note that most credible interval width results for $M1$ are so large that they are outside the limits of the plot.) With regards to the COVID-19 pandemic, this emphasizes the importance of conducting some amount of "unbiased testing," even if the sample sizes are relatively small; see Cochran (2020).

Finally, note that, if $k' = 0$, mixing can be problematic if λ is small. Indeed, for the $M1$ model, convergence/mixing failures occurred in about 1% of simulation runs, when $\lambda = 0.05$, and occurred in less than 0.004% of simulation runs when $\lambda > 0.05$. With very small λ values (e.g., $\lambda < 0.001$), we suspect that convergence may simply be impossible. This is no doubt due to the identifiability issues discussed in Section 2.2 and in the Appendix (Section A.2). If $k' = 0$, the model benefits greatly (in terms of mixing and identifiability) from specifying more informative priors.

6. Application: IFR of COVID-19 in Europe. Reducing uncertainty around the severity of COVID-19 was of great importance to policy makers and the public during the early stages of the pandemic and continues to be a top priority (Ioannidis (2020b), Lipsitch (2020)). Comparisons between the COVID-19 and seasonal influenza IFRs impacted the timing and degree of social distancing measures and highlighted the need for more accurate estimates for the severity of both viruses (Faust (2020)). A lack of clarity means that policy makers are unsure if cross-population differences are related to clinically relevant heterogeneity (i.e., due to large τ) or to spurious heterogeneity driven by testing and reporting biases (i.e., due to large γ).

We demonstrate how the proposed model could be used to estimate the IFR of COVID-19 in Europe during the spring of 2020. Note that the main purpose of this analysis is to demonstrate the feasibility of the proposed model. As such, we keep things relatively simple. For instance, we only consider countries belonging to the E.U./E.E.A. (European Economic Area), the United Kingdom and Switzerland, as these could be considered a reasonably homogeneous group. However, we exclude Belgium since, uniquely, the country counts all suspect deaths in nursing homes as COVID-19 deaths (Lee (2020)).

We selected $k' = 5$ studies for which we assume there is no preferential testing. To do so, we considered all European seroprevalence studies reporting an IR estimate (along with a 95% confidence/credible interval) listed in the systematic review by Ioannidis (2020a). From these we selected only those studies that claimed to achieve a representative or random sample from their study population.

It is important to note that the seroprevalence studies were conducted amongst populations which were particularly hard hit by infection. The result is that these populations are not necessarily representative of the overall European population. It is unclear how this might impact our model estimates. Also, while some of the seroprevalence studies report the exact number of tests conducted (T) and the number of confirmed cases recorded (CC), to obtain estimates for the infection rate there are numerous adjustments (e.g., adjusting for testing sensitivity and specificity). Rather than work with the raw T_k and CC_k numbers published in the seroprevalence studies, we calculate effective data values for CC_k and T_k based on a binomial distribution that corresponds to the reported 95% CI for the IR. By "inverting binomial confidence intervals" in this way, we are able to properly use the adjusted numbers for each of the five seroprevalence studies. This is a similar approach to the strategy employed

by Kümmerer, Berens and Macke (2020) who assume that the IR follows a Beta distribution with parameters chosen to match the 95% CI published in Streeck et al. (2020). In the Supplementary Material (Campbell et al. (2022)) we go over the seroprevalence study data in detail.

We obtained national official COVID-19 statistics as reported by Our World in Data (OWID (2020)). Complete data was available for 26 countries which brings the total number of groups to $K = 31$. The CC_k and T_k numbers were selected, as reported on May 1, 2020 (or the earliest date during the following week for which data was available). Numbers for D_k for $k = 6, \ldots, K$ were obtained from 14 days afterward to allow for the known delay between the onset of infection and death, taking into consideration the delay between the onset of infection and the development of detectable antibodies; see Wu et al. (2020) and Linton et al. (2020).

Note that our T_k numbers are not ideal, since some countries report the number of people tested, while others report the total number of tests (which will be higher if a single person is tested several times). Also note that, as stated in Section 2.1, the K different groups should, in principle, be entirely independent samples. This is clearly not the case with the European data (case in point, there are three different groups from within Switzerland; $k = 2$, $k = 5$ and $k = 30$).

We included several covariates about each country's population to explain variation in IR and IFR. Specifically, for the IR we consider: (1) the number of days since the country reported 10 or more confirmed infections ("Days since outbreak") (as reported by Hale et al. (2020)), (2) the number of days between a country's first reported infection and the imposition of social distancing measures ("Days until lockdown") (calculated based on when the Government Response Stringency Index (GRSI) reached 20 or higher, as reported in OWID (2020)) and (3) the population density ("Pop. density") (as reported by OWID (2020) and other publicly available sources[2]). For the IFR we consider: (1) the share of the population that is 70 years and older ("Prop. above 70 y.o.") (as reported in Ioannidis (2020a) or OWID (2020)) and (2) the number of hospital beds per 1000 people ("Hosp. beds per 1000").[3] Tables 1 and 2 in the Supplementary Material (Campbell et al. (2022)) list all the data used in the analysis.

6.1. *Using only seroprevalence studies.* Using only the seroprevalence studies (i.e., only the first $k' = 5$ studies listed in Table 1 in the Supplementary Material (Campbell et al. (2022))), we fit the model as described in Section 4.1 (with $\eta = 0.1$) without any adjustment for covariates. (With only $K = 5$ groups, there are few degrees of freedom available for including group-level covariates.) The model was fit using Stan (Carpenter et al. (2017)) with four independent chains, each with 20,000 draws (10% burn-in, thinning of five). Figure 2 plots the posterior medians obtained for the IR_k and IFR_k parameters (for $k = 1, \ldots, 5$) with 95% HPD CIs. We also plot, in black, the posterior median of $\mathrm{g}^{-1}(\beta)$ and $\mathrm{g}^{-1}(\theta)$ ("Overall"). Our estimate for the overall IFR is $\mathrm{g}^{-1}(\theta) = 0.54\%$, 95% C.I. $= [0.43\%, 0.68\%]$. We note that the five IFR_k estimates obtained are very homogeneous. This is no doubt partly due to the punitive nature of our prior on τ (i.e., due to setting $\eta = 0.1$). Indeed, when the model is fit with $\eta = 1$, the IFR_k estimates are more heterogeneous (the posterior median estimates obtained with $\eta = 1$ are $\mathrm{IFR}_k = 0.50, 0.48, 0.65, 0.53, 0.60$, for $k = 1, \ldots, 5$, respectively, and $\mathrm{g}^{-1}(\theta) = 0.54\%$, 95% C.I. $= [0.37\%, 0.80\%]$).

[2]For Geneva (https://www.bfs.admin.ch/bfs/en/home/statistics/regional-statistics/regional-portraits-key-figures/cantons/geneva.html); for Gangelt (https://en.wikipedia.org/wiki/Gangelt); for Split-Dalmatia (https://en.wikipedia.org/wiki/Split-Dalmatia_County); for Zurich (https://www.bfs.admin.ch/bfs/en/home/statistics/regional-statistics/regional-portraits-key-figures/cantons/zurich.html).

[3]Obtained from OWID (2020) or from www.bfs.admin.ch for Geneva and Zurich cantons.

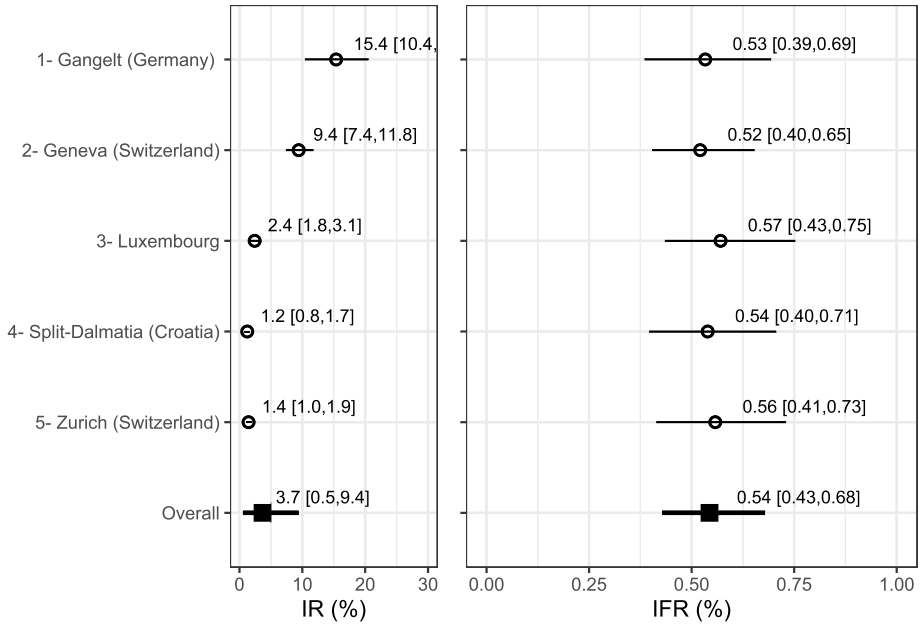

FIG. 2. *Posterior median estimates for the IR_k and IFR_k variables (for $k = 1, \ldots, 5$) with 95% HPD CIs. Also plotted, under the label "Overall," is the posterior median estimate and 95% HDP CI of $g^{-1}(\beta)$ and $g^{-1}(\theta)$. These results correspond to the large-P model with $\eta = 0.1$ which pools information from five seroprevalence studies ($K = 5$ and $k' = 5$).*

6.2. *Using all the data.* We fit the model, as described in Section 4.1, to all the data (listed in Tables 1 and 2 in the Supplementary Material (Campbell et al. (2022))) with $k' = 5$, $K = 31$, $h = 3$ and $q = 2$. Covariates were defined as the centered and scaled logarithm of each metric as follows:

$$X_{[1]} = \text{center-scale}\big(\log(\text{“Days since outbreak”})\big);$$
$$X_{[2]} = \text{center-scale}\big(\log(\text{“Days until lockdown” + 1})\big);$$
$$X_{[3]} = \text{center-scale}\big(\log(\text{“Population density”})\big);$$
$$Z_{[1]} = \text{center-scale}\big(\log(\text{“Prop. above 70 y.o.”})\big), \quad \text{and}$$
$$Z_{[2]} = \text{center-scale}\big(\log(\text{“Hosp. beds per 1000”})\big).$$

Standard normal priors ($\mathcal{N}(0, 1)$) were used for each of β_1, β_2, β_3, θ_1 and θ_2. All other priors were defined, as in Section 3, with $\eta = 0.1$ and $\lambda = 0.05$. The model was fit using Stan (Carpenter et al. (2017)) with four independent chains, each with 20,000 draws (10% burn-in, thinning of five).

Figure 3 plots the estimates (posterior medians) obtained for the IR_k and IFR_k variables (for $k = 1, \ldots, 31$) with 95% HPD CIs. We also plot the posterior medians of $g^{-1}(\beta)$ and $g^{-1}(\theta)$ ("Overall"). Our estimate for the overall IFR is $g^{-1}(\theta) = 0.53\%$, 95% C.I. = [0.38%, 0.70%].

Table 1 lists posterior medians with HPD 95% CIs for the main parameters of interest. The positive values for $\beta_1(0.21, 95\%\text{CI} = [-0.10, 0.52])$ and β_2 (0.45, 95% CI = [0.15, 0.78]) suggest that the IR increases with increasing time since the initial disease outbreak and with increasing time between the first reported infection and the imposition of social distancing measures. The positive value for β_3 (0.76, 95% CI = [0.47, 1.04]) suggests that a higher population density is associated with a higher IR. The negative value for θ_2 (-0.43, 95% CI = $[-0.62, -0.24]$) suggests that countries with fewer hospital beds have higher IFRs.

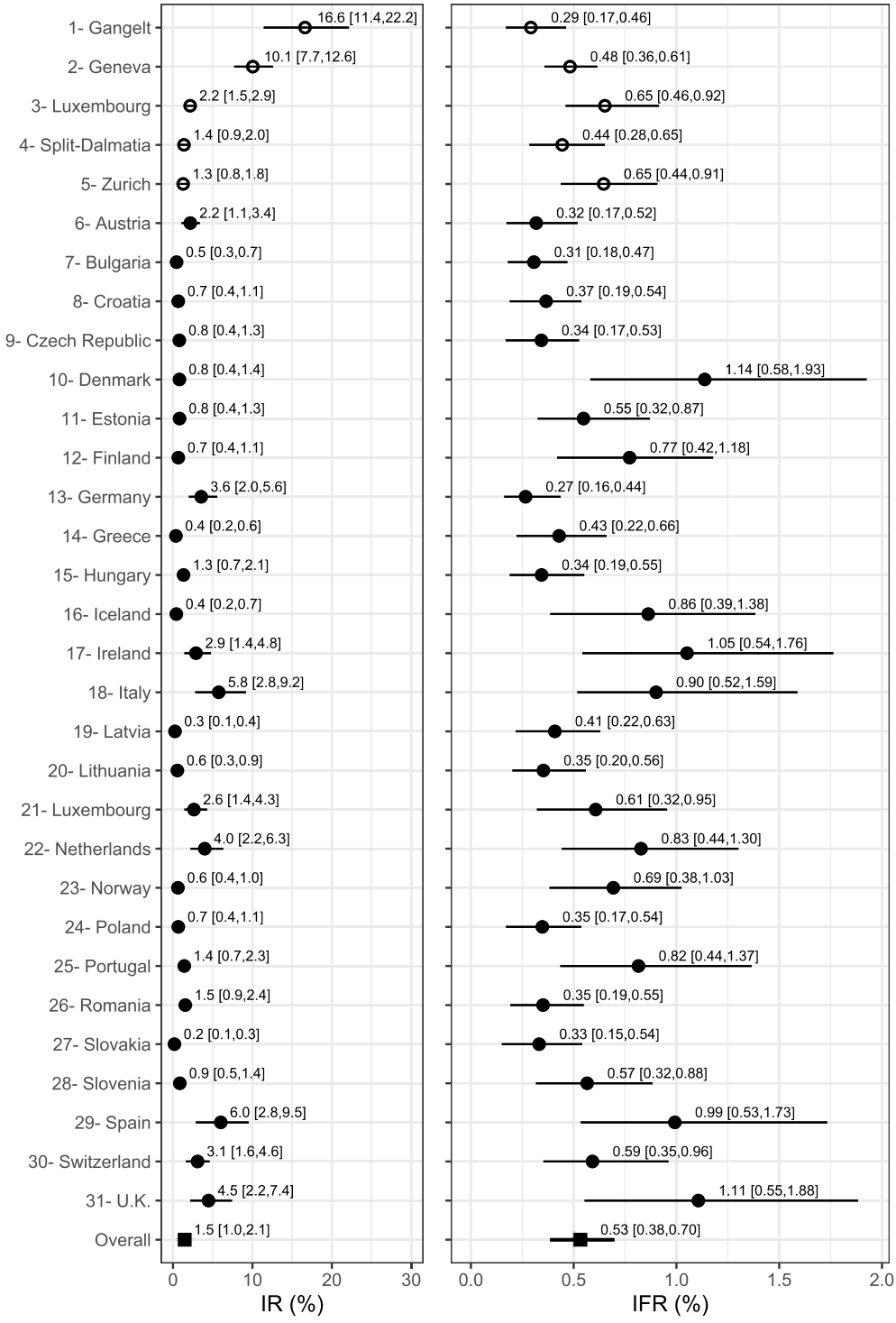

FIG. 3. *Posterior median estimates for the IR_k and IFR_k variables (for $k = 1, \ldots, 31$) with 95% HPD CIs. Also plotted, under the label "Overall," is the posterior median estimate and 95% HDP CI of $g^{-1}(\beta)$ and $g^{-1}(\theta)$. These results correspond to the large-P model with $\eta = 0.1$ which pools information from five seroprevalence studies and data from nationally reported statistics for 26 European countries ($K = 31$ and $k' = 5$).*

TABLE 1

Posterior parameter estimates (posterior medians and 95% HPD CIs) from large-P model fit to data from only the seroprevalence studies (left) and from the full dataset (right). The large-P model is fit with priors specified by $\lambda = 0.05$ and $\eta = 0.1$

	Seroprevalence data		All data	
	Estimate	95% CI	Estimate	95% CI
$g^{-1}(\theta) \times 100$	0.544	[0.428, 0.679]	0.533	[0.385, 0.698]
$g^{-1}(\beta) \times 100$	3.677	[0.548, 9.432]	1.473	[0.968, 2.053]
θ	−5.212	[−5.441, −4.980]	−5.232	[−5.526, −4.936]
β	−3.285	[−4.479, −2.117]	−4.210	[−4.585, −3.845]
θ_1 ("Prop. above 70 y.o.")			−0.005	[−0.155, 0.171]
θ_2 ("Hosp. beds per 1000")			−0.430	[−0.622, −0.238]
β_1 ("Days since outbreak")			0.208	[−0.103, 0.520]
β_2 ("Days until lockdown")			0.460	[0.154, 0.783]
β_3 ("Pop. density")			0.756	[0.474, 1.043]
τ	0.078	[0.001, 0.203]	0.196	[0.021, 0.343]
σ	1.171	[0.604, 1.967]	0.688	[0.499, 0.917]
γ			9.141	[4.881, 14.055]

We were quite surprised to see that our estimate of θ_1 (-0.01, 95% CI $= [-0.16, 0.17]$) was not decidedly larger and positive, given that age is known to be an important predictor of COVID-19 complications and death (O'Driscoll et al. (2020)). There are several reasons why we might have obtained this result. First, statistical power may have been compromised by insufficient heterogeneity in the age-structure across different countries, as captured by the proportion aged over 70 metric. Furthermore, as with a standard multivariable regression of observational data, the estimate of θ_1 may suffer from bias, due to unobserved confounding and/or multicollinearity.

Note that the overall IFR estimate and its uncertainty are very similar whether or not the data from nationally reported statistics are included in the analysis (only sero-studies: $g^{-1}(\theta) = 0.54\%$, 95% C.I. $= [0.43\%, 0.68\%]$ *vs.* all data: $g^{-1}(\theta) = 0.53\%$, 95% C.I. $= [0.38\%, 0.70\%]$). One might, therefore, question what added value the expanded analysis provides. We have two comments on this point.

First, since we did not incorporate any substantial prior information about the magnitude of preferential testing into the model (i.e., we selected very "weakly informative" priors for the ϕ parameters), we should not expect the point estimate of θ to differ substantially between the two analyses.

Second, even if the estimate for the overall IFR is left unchanged, incorporating the additional data into an expanded analysis is still worthwhile. Simulation study results (see Section 5.2) show that a considerable sharpening of information is possible in many scenarios. Without actually implementing the expanded analysis, it would be impossible to know whether or not such a sharpening would occur in the present context. Moreover, unlike the analysis which uses only seroprevalence study data, the expanded analysis allows one to obtain valuable country-specific IFR and IR estimates as well as to obtain estimates for the association between the IFR/IR and a number of different explanatory factors.

The model can no doubt be improved by using appropriately specified informed priors for the ϕ parameters, based on what is known about COVID-19 testing in different countries. For example, in related work, Grewelle and De Leo (2020) assume that testing capacity is directly proportional to the case load in each country (where testing capacity is estimated by

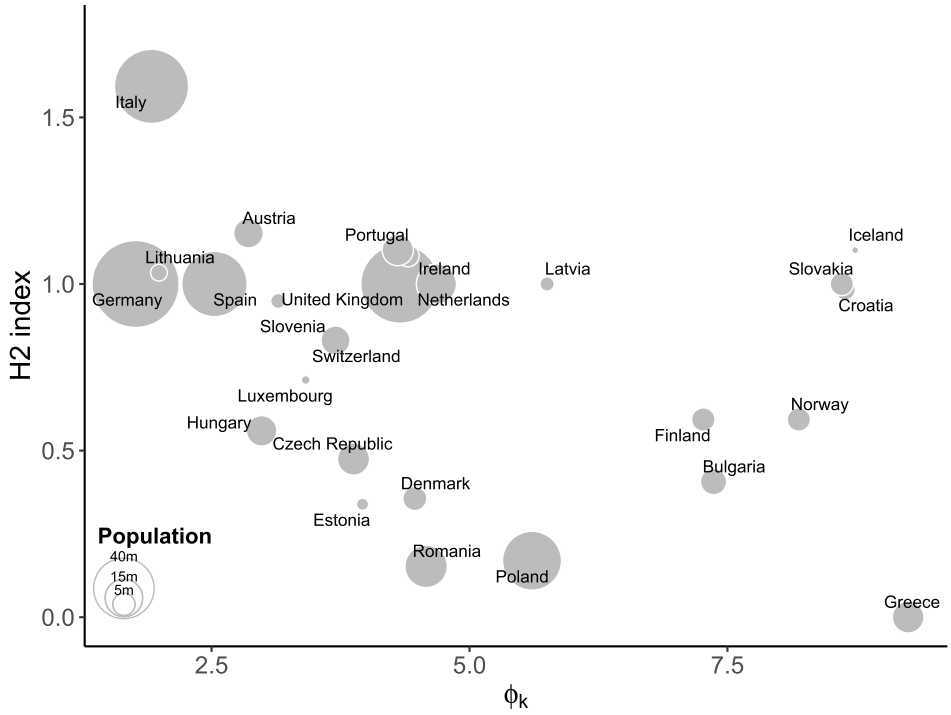

FIG. 4. *Scatter plot shows of the average H2 index for each country (for the period between February* 1, 2020 *and April* 1, 2020) *vs. the posterior median ϕ_k value. Circle size corresponds to population (P_k). These results correspond to the large-P model with $\eta = 0.1$ which pools information from five seroprevalence studies and data from nationally reported statistics for* 26 *European countries.*

tests performed per positive case[4]). As another example, the "H2 index" (Hale et al. (2020)), which purportedly reflects official government policy on who has access to testing within a given country, could also be used to define informed priors for the ϕ parameters in a more sophisticated version of our model.

We were curious as to whether the model estimates (posterior medians) we obtained for ϕ_k (for $k = 6, \ldots, 31$) might be predictive of the H2 index. Using the data made available by Hale et al. (2020), we calculated the average H2 index for each country in our analysis, for the period between February 1, 2020 and April 1, 2020. Roughly speaking, a high H2 value indicates broad access to testing (i.e., available to the general public), whereas a low H2 value reflects a testing policy that restricts testing to only those who have symptoms or meet specific criteria. Thus, countries with high H2 values should, in theory, have small values of ϕ_k and vice versa. That prediction is generally supported by the results seen in Figure 4, although Iceland, Slovakia and Croatia appear to be exceptions.

7. Discussion.

7.1. *Model limitations.* Estimation of the IFR is very challenging, due to the fact that it is a ratio of numbers where both the numerator and denominator are subject to a wide range of biases. Our proposed model seeks to address only one particular type of bias pertaining

[4]Grewelle and De Leo (2020) are, thereby, able to infer the "global IFR" using simple weighted linear regression (i.e., regressing $\log(D_k/CC_k) \sim \log(IFR_k) + \beta_1(CC_k/T_k)$, for $k = 1, \ldots, K$, where β_1 is an unknown nuisance parameter).

to the denominator, the bias in the number of cases due to preferential testing. With this in mind, we wish to call attention to several other important sources of bias.

Cause of death information may be very inaccurate. To overcome this issue, many suggest looking to "excess deaths" by comparing aggregate data for all-cause deaths from the time during the pandemic to the years prior (Leon et al. (2020)). Using this approach and a simple Bayesian binomial model, Rinaldi and Paradisi (2020) are able to obtain IFR estimates without relying on official (possibly inaccurate) data for the number of COVID-19 deaths.

Some people who are currently sick will eventually die of the disease but have not died yet. Due to the delay between disease onset and death, the number of confirmed and reported COVID-19 deaths at a certain point in time will not reflect the total number of deaths that will occur among those already infected (right-censoring). This will result in the number of recorded deaths underestimating the true risk of death. The denominator of the IFR must be the number of cases *with known outcomes*.

The model assumes that no individuals are tested more than once. This is an important practical limitation. In addition, the model, as currently proposed, fails to account for the (unknown) number of false positive and false negative tests. When both the test specificity and the infection rate is low, false positives can substantially inflate the estimated infection rate and, as a consequence, the IFR could be biased downward. In principle, the model could accommodate for this by specifying priors for test sensitivity and specificity; see Kümmerer, Berens and Macke (2020), Gelman and Carpenter (2020) and Neil et al. (2020).

Finally, because the model uses data that are aggregated at the group level, estimates are potentially subject to ecological bias (Pearce (2000)). While including group-level covariates may help reduce variability in the estimates, adjustment using group-level covariates can also lead to misleading results (Li and Hua (2020)).

7.2. *Concluding remarks*. Obtaining representative data remains challenging and costly. Efforts to better understand the distribution of SARS-CoV-2 infection (and its lethality) at the population level have, unfortunately, been met by recruiting challenges (Gudbjartsson et al. (2020), Bendavid et al. (2020)), leading to an overrepresentation of people who are concerned about their exposure and/or an underrepresentation of individuals who are selfquarantining, isolating or hospitalized because of the virus. In the absence of large-scale unbiased data, researchers must work with whatever data is available. Our model suggests a coherent and feasible way to do just this.

We demonstrated our proposed model with an application to European COVID-19 data in which we relied on data from seroprevalence studies that selfreported as representative. When combined with data from nationally reported statistics, this data enabled us to obtain appropriate estimates not only for the overall IFR but also for country-specific IFRs and IRs as well as for the association between these and various explanatory factors.

We note that our estimate for the overall IFR in Europe (0.53%, 95% C.I. = [0.38%, 0.70%]) is somewhat lower than an estimate obtained by the meta-analysis of Meyerowitz-Katz and Merone (2020) from European seroprevalence studies (IFR = 0.77%, 95% C.I. = [0.55%, 0.99%]) and reiterate that the primary intention of our analysis was to demonstrate the feasibility of the proposed model. That being said, we selected the five seroprevalence studies based on the review of Ioannidis (2020a) and suspect that the differing inclusion/exclusion criteria between Ioannidis (2020a) and Meyerowitz-Katz and Merone (2020) are the main reason why our estimates are not more similar. A recent large-scale analysis for Spain, based on a nationwide population-based seroprevalence study (Pollán et al. (2020)), concludes that, for the Spanish population, "overall infection fatality risk was 0.8% [...] for confirmed COVID-19 deaths and 1.1% [...] for excess deaths [...]" (Pastor-Barriuso et al. (2020)). This is in reasonable agreement with our estimate for Spain of 0.99% (95% credible interval of [0.55%, 1.73%]).

In a similar study, O'Driscoll et al. (2020) conduct a Bayesian analysis using data from 22 national-level seroprevalence surveys and sex and age-specific COVID-19-associated death data from 45 countries. Their data is much more granular and their model specifies many more detailed assumptions. For example, O'Driscoll et al. (2020) assume "a gamma-distributed delay between onset [of infection] and death" and assume different risks of infection for "individuals aged 65 years and older, relative to those under 65," since "older individuals have fewer social contacts and are more likely to be isolated through shielding programmes." O'Driscoll et al. (2020) conclude that "population-weighted IFR estimates by the ensemble model are highest for countries with older populations such as [...] Italy (0.94%; 95% credible interval, 0.80–1.08%)." While our model did not identify a significant association between IFR and countries with older populations (much to our surprise), many of our country-specific IFR estimates are quite similar to those reported in O'Driscoll et al. (2020). For example, we obtained an IFR for Italy of 0.90% (95% credible interval of [0.52%, 1.59%]).

In the Introduction, we identified two important questions. First (Q1), is it possible to reliably estimate the IFR without any information about the degree to which the data are biased by preferential testing? And second (Q2), when representative samples are available, can samples with an unknown degree of preferential testing contribute valuable information? The proposed Bayesian model suggests that reliable estimation of the IFR at the group level is indeed possible, to a certain extent, when existing data do not arise from a random sample from the target population. Importantly, the key to (partial) identifiability is sufficient heterogeneity in the degree of preferential testing across groups and sufficient homogeneity in the group-specific IFR. We also saw that combining both types of data (biased and unbiased data) can be superior to ignoring data that may be skewed by preferential testing. When fit with an appropriate model, biased data can supplement available representative data in order to refine one's inference and/or shed light on the impact of explanatory factors.

In a typical situation of drawing inference from a single biased sample, obtaining appropriate estimates is challenging, if not impossible, without some sort of external validation data. Intuition suggests that one might only be able to do a sensitivity analysis with respect to the impact of bias. Indeed, applying prior distributions for the degree of preferential testing and proceeding with Bayesian inference could be regarded as a probabilistic form of sensitivity analysis (Greenland (2005)). What is perhaps less intuitive, and what we demonstrated with the proposed model, is that if one has multiple samples of biased data, each subject to a different degree of bias, the "heterogeneity of bias" can help inform what overall adjustment is required for appropriate inference.

The aggregation of data from both biased and unbiased samples is a problem that applies to many types of evidence synthesis (and is often overlooked) (De Angelis et al. (2015), Birrell, De Angelis and Presanis (2018)). In that sense, the solutions we put forward may be more broadly applicable. Future work should investigate whether the "heterogeneity of bias" principle (see Section A.2) can be used to derive appropriate estimates in a meta-analysis where individual studies are subject to varying degrees of bias, due to unobserved confounding or measurement error (e.g., Campbell et al. (2020)).

APPENDIX

A.1. Details for the noncentral hypergeometric distribution. In Section 2.1, we consider the distribution of CC|C as following Wallenius' noncentral hypergeometric (NCHG) distribution such that

$$\mathrm{CC}|\mathrm{C} \sim \mathrm{NCHG}(\mathrm{C}, \mathrm{P} - \mathrm{C}, \mathrm{T}, \phi), \tag{12}$$

TABLE 2
Illustrative example data with varying degrees of preferential sampling: $\gamma = 0$, $\gamma = 4$, $\gamma = 11$ *and* $\gamma = 22$

	Observed			$\gamma = 0$	4	11	22	Unobserved			$\gamma = 4$	11	22
k	P_k	T_k	D_k	CC_k	CC_k	CC_k	CC_k	C_k	IR_k	IFR_k	ϕ_k	ϕ_k	ϕ_k
1	3061	190	11	24	21	32	27	430	0.140	0.018	1	1	1
2	482	43	2	15	11	12	24	99	0.206	0.020	1.36	2	3
3	1882	101	20	32	40	55	74	570	0.303	0.022	1.73	3	5
4	1016	67	2	14	24	33	38	193	0.190	0.017	2.09	4	7
5	1269	109	4	13	34	54	67	201	0.159	0.021	2.45	5	9
6	3670	276	9	53	70	140	162	484	0.132	0.021	2.82	6	11
7	2409	139	7	17	34	70	94	329	0.137	0.019	3.18	7	13
8	1074	81	13	42	65	68	77	565	0.526	0.019	3.55	8	15
9	3868	289	16	60	142	205	247	821	0.212	0.019	3.91	9	17
10	151	13	2	1	5	11	8	24	0.160	0.019	4.27	10	19
11	430	25	1	6	9	16	18	70	0.164	0.019	4.64	11	21
12	429	40	2	11	23	31	33	105	0.245	0.019	5	12	23

where the degree of preferential testing corresponds to the ϕ noncentrality parameter. The probability mass function of this distribution (Lyons (1980), Fog (2008)) can be written out as

$$f(x) = \binom{C}{x}\binom{P-C}{T-x}\int_0^1 \left(1 - t^{\phi_k/B}\right)^x \left(1 - t^{1/B}\right)^{T-x} \mathrm{d}t, \tag{13}$$

where $B = \omega(C - x) + (P - C - (T - x))$. Note that when the noncentrality parameter, ϕ, equals 1, the distribution is equivalent to the standard central hypergeometric distribution.

Wallenius' noncentral hypergeometric is often confused with Fisher's noncentral hypergeometric distribution. Wallenius' noncentral hypergeometric distribution describes the situation where a predetermined number of items are seleted one by one, whereas Fisher's noncentral hypergeometric distribution describes a situation where the total number of items drawn is only known after the experiment.

A.2. Issues of (un)identifiability. Table 2 provides a small artificial dataset to help illustrate the type of data being described and the impact of different degrees of preferential testing. In this dataset we have $K = 12$ groups and the (unknown) infection rate varies substantially from 13% to 53%. The unknown infection fatality rate only varies slightly, from 0.017% to 0.022%. Values for ϕ_k in this dataset are evenly distributed between 1 and $\gamma + 1$, for four different values of $\gamma = 0$, 4, 11 and 22. When $\gamma = 0$, the number of true cases (i.e., actual infections) is approximately 14 times higher than the number of confirmed cases. In contrast, when $\gamma = 22$, the number of true cases is only about five times higher than the number of confirmed cases.

Here, we present an asymptotic argument which lays bare the flow of information. Consider a situation in which an infinite amount data are available. In so-called "asymptotia," we have that populations are approaching infinite size (i.e., for $k = 1, \ldots, K$, we have $P_k \to \infty$) and that the number of tests also approaches infinity (i.e., for $k = 1, \ldots, K$, we have $T_k \to \infty$). Recall that a hypergeometric distribution is asymptotically equivalent to a binomial distribution. As such, we consider the following approximation (as in Section 4.1):

$$D_k \sim \text{Binom}(P_k, a_k); \quad \text{and} \quad CC_k \sim \text{Binom}(T_k, b_k),$$

where $a_k = IFR_k \times IR_k$ and $b_k = 1 - (1 - IR_k)^{\phi_k}$.

Presume that the a priori defensible information about the preferential sampling in the kth group is expressed in the form

$$\phi_k \in [\underline{\phi}_k, \bar{\phi}_k], \tag{14}$$

that is, $\underline{\phi}_k$ and $\bar{\phi}_k$ are investigator-specified bounds on the degree of preferential sampling for that jurisdiction. If one is certain that cases are as likely, or at least as likely, to be tested as noncases, $\underline{\phi}_k = 1$ is appropriate. If testing is known to be entirely random for the kth group, one would set $\underline{\phi}_k = \bar{\phi}_k = 1$.

Note that, for fixed (a_k, b_k), IFR_k is a function of ϕ_k with the form

$$\text{IFR}_k(\phi_k) = \frac{a_k}{1 - (1 - b_k)^{(1/\phi_k)}}. \tag{15}$$

Examining (15), knowledge of (a_k, b_k), in tandem with (14), restricts the set of possible values for IFR_k. In fact, it is easy to verify that (15) is monotone in ϕ_k; hence the restricted set is an interval. We write this interval as $I_k(a_k, b_k, \underline{\phi}_k, \bar{\phi}_k)$ or simply as I_k for brevity. This is the jurisdiction-specific *identification interval* for IFR_k. As we approach asymptotia for the kth group, all values inside the interval remain plausible, while all values outside are ruled out; see Manski (2003). This is the essence of the *partial identification* inherent to this problem.

Thinking now about the meta-analytic task of combining information, we envision that both ϕ_k and IR_k could exhibit considerable variation across jurisdictions. However, the variation in IFR could be small, particularly if sufficient jurisdiction-specific covariates are included (see Section 4.2). That is, after adjustment for a jurisdiction's age-distribution, healthcare capacity and so on, residual variation in IFR could be very modest. When modeling, we would invoke such an assumption via a prior distribution. For understanding in asymptotia, however, we simply consider the impact of an a priori bound on the variability in IFR. Let τ be the standard deviation of IFR across jurisdictions. Then, we presume τ does not exceed an investigator-specified upper bound of $\bar{\tau}$, that is, $\tau \leq \bar{\tau}$.

The jurisdiction-specific prior bounds on the extent of preferential sampling and the prior bound on IFR variation across jurisdictions, along with the limiting signal from the data in the form of (a, b), gives rise to an identification region for the average infection fatality rate, $\overline{\text{IFR}} = K^{-1} \sum_{k=1}^{K} \text{IFR}_k$. Formally, this interval is defined as

$$I(a, b, \underline{\phi}, \bar{\phi}, \bar{\tau}) = \left\{\overline{\text{IFR}} : \tau \leq \bar{\tau}, \text{IFR}_k \in I_k(a_k, b_k, \underline{\phi}_k, \bar{\phi}_k), \forall k \in \{1, \ldots, K\}\right\}. \tag{16}$$

Again, the interpretation is direct: in the asymptotic limit, all values of $\overline{\text{IFR}}$ inside this interval are compatible with the observed data, and all values outside are not. The primary question of interest is whether this interval is narrow or wide under realistic scenarios, since this governs the extent to which we can learn about $\overline{\text{IFR}}$ from the data.

In general, evaluating (16) for given inputs is an exercise in quadratic programming nested within a grid search, hence can be handled with standard numerical optimisation. However, the special case of $\bar{\tau} = 0$ is noteworthy in terms of developing both scientific and mathematical intuition. Consequently, we explore this case in some depth in what follows.

Scientifically, $\bar{\tau} = 0$ represents the extreme limit of an a priori assumption that, possibly after covariate adjustment, IFR is a 'biological constant' which does not vary across jurisdictions. If the prospects for inference are not good when this assumption holds, they will be even less good under the less strict assumption that the IFR heterogeneity is small, but not necessarily zero. Mathematically, the case is much simpler, with (16) reducing to

$$I(a, b, \underline{\phi}, \bar{\phi}, 0) = \bigcap_k I_k(a_k, b_k, \underline{\phi}_k, \bar{\phi}_k). \tag{17}$$

As intuition must have it, without heterogeneity a putative value for the "global" IFR is compatible with the observed data if and only if it is compatible with the data from *every* jurisdiction individually.

To illustrate, consider a scenario with $K = 12$ jurisdictions, with a constant infection fatality rate of 2%, that is, $\text{IFR}_k = 0.02$, for $k = 1, \ldots, 12$. Say that the infection rates for these jurisdictions lie between 0.132 and 0.526, as per Table 1. Furthermore, say that the unknown ϕ_k values range between 1 and 23, as per the rightmost column ($\gamma = 22$) of Table 1.

Now, say the investigator pre-pecifies $(\underline{\phi}_k, \bar{\phi}_k) = (1, 40)$ for all k. As such, the a priori bounds are correct for all jurisdictions. The resulting jurisdiction-specific identification intervals, I_k, are depicted in the bottom left-hand panel of Figure 5. (The top and middle left-hand panels correspond to the identical situation but with ϕ_k values listed in the $\gamma = 4$ and $\gamma = 11$ columns of Table 1, respectively.) Also depicted by the grey rectangle is the global identification interval, that is, the intersection of the individual intervals. In the present scenario

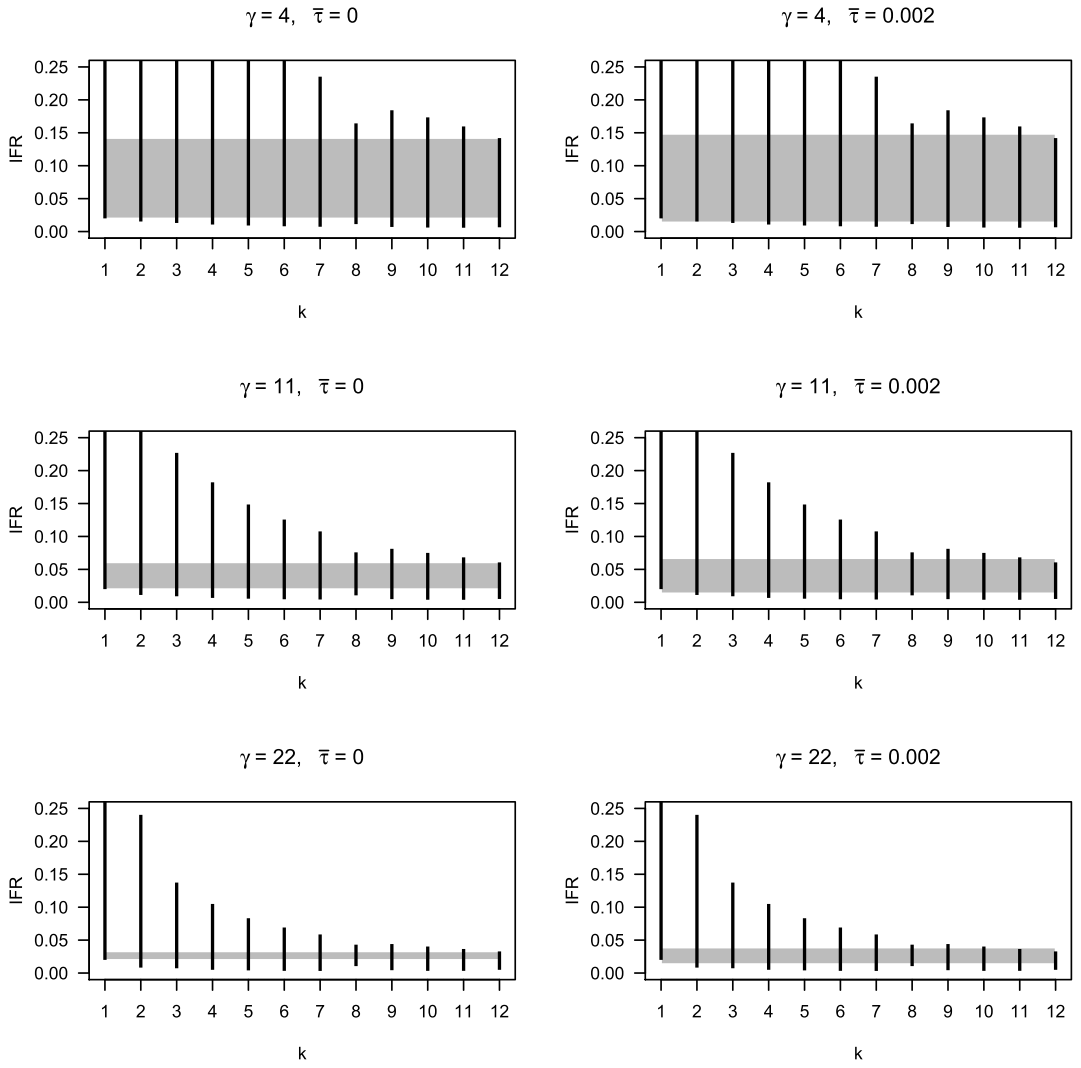

FIG. 5. *Consider the meta-analytic task of combining information in a scenario with $K = 12$ jurisdictions. Here, the black lines correspond to jurisdiction-specific identification intervals for the IFR, and the grey rectangles corresponds to the global identification interval for IFR. Left-hand panels correspond to assumption of $\bar{\tau} = 0$ such that the global identification interval is simply the intersection of the individual intervals. Right-hand panels correspond to $\bar{\tau} = 0.002$.*

($\gamma = 22$), this is indeed narrow, ranging from 0.0200 to 0.0328. (For the $\gamma = 4$, $\gamma = 11$ and $\gamma = 22$ scenarios, the global identification intervals are [0.0200, 0.1419], [0.0200, 0.0606] and [0.0200, 0.0328], respectively.) Thus, depending on the range in ϕ_k values, that is, depending on the "heterogeneity of bias," it appears that data can contribute substantial information about the (constant) infection fatality rate.

As can be seen immediately from Figure 5 (left-hand panels), in the present example the binding constraints arise from the first and twelfth jurisdictions which happen to have the least and most amounts of preferential testing. However, this pattern does not hold in general. One can easily construct pairs of infection rates for which the jurisdiction with more preferential testing has a smaller upper endpoint for I_k and/or a larger lower endpoint. Thus, the values of ϕ_k alone do not determine which two jurisdictions will provide the binding information about $\overline{\text{IFR}}$.

Figure 5 (right-hand panels) shows how the global identification interval is wider when $\overline{\tau} = 0.002$. For reference, for the IFR values listed in Table 1, $\tau = \text{SD}(\text{IFR}_{1:12}) = 0.00124$. For the $\gamma = 4$, $\gamma = 11$ and $\gamma = 22$ scenarios, the global identification intervals outlined by the grey rectangles are [0.0139, 0.1483], [0.0137, 0.0670] and [0.0137, 0.0386], respectively.

Now, consider the evaluation of (16) for $\overline{\tau} > 0$, that is, where a limited heterogeneity in IFR is permitted. Recall that quadratic programming constitutes the minimization of a quadratic function subject to linear constraints, and these may be a mix of equality and inequality constraints. Let x be a candidate value, which we will test for membership in the identification interval. To perform this test, we use a standard quadratic programming package (quadprog) to minimize the quadratic function Var(IFR), subject to the equality constraint $\overline{\text{IFR}} = x$ and the $2K$ inequality constraints which restrict IFR_k to the interval I_k for each k. By the definition of (6), then x belongs in the identification interval if and only if the minimized variance does not exceed $\overline{\tau}^2$.

Thus, a simple grid search over values of x numerically determines the identification interval. Note that so long as a and b arise from values of ϕ within the prescribed bounds, the underlying value of $\overline{\text{IFR}}$ must belong to the identification interval. Thus, two numerical searches can be undertaken. One starts at the underlying value and tests successively larger x until a failing value is obtained. The other starts at the underlying value and does the same, but moving downward.

Acknowledgments. We wish to thank Joe Watson for his input early on and expertise on preferential sampling.

Funding. This work was supported by the European Union's Horizon 2020 research and innovation programme under ReCoDID grant agreement No. 825746 and by the Canadian Institutes of Health Research, Institute of Genetics (CIHR-IG) under Grant Agreement No. 01886-000.

SUPPLEMENTARY MATERIAL

Additional model details, data, and analysis code (DOI: 10.1214/21-AOAS1499SUPP; .zip). The Supplemental Material zip folder contains a document describing: (1) the nimble MCMC details, (2) the seroprevalence study data, and (3) the data tables. The folder also includes all R code used for the Europe data analysis along with the three required datasets and their corresponding data dictionaries.

REFERENCES

ANDERSON, R. M., HEESTERBEEK, H., KLINKENBERG, D. and HOLLINGSWORTH, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **395** 931–934.

BENDAVID, E., MULANEY, B., SOOD, N., SHAH, S., LING, E., BROMLEY-DULFANO, R., LAI, C., WEISSBERG, Z., SAAVEDRA, R. et al. (2020). COVID-19 antibody seroprevalence in Santa Clara County, California. medRxiv.

BERGER, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer, Berlin.

BIRRELL, P. J., DE ANGELIS, D. and PRESANIS, A. M. (2018). Evidence synthesis for stochastic epidemic models. *Statist. Sci.* **33** 34–43. MR3757502 https://doi.org/10.1214/17-STS631

BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662 https://doi.org/10.2307/1390675

CAMPBELL, H., DE JONG, V. M., MAXWELL, L., DEBRAY, T., JAENISCH, T. and GUSTAFSON, P. (2020). Measurement error in meta-analysis (MEMA)–a Bayesian framework for continuous outcome data. *Res. Synth. Methods*. https://doi.org/10.1002/jrsm.1515.

CAMPBELL, H., DE VALPINE, P., MAXWELL, L., DE JONG, V. M., DEBRAY, T. P., JAENISCH, T. and GUSTAFWSON, P. (2022). Supplement to "Bayesian adjustment for preferential testing in estimating infection fatality rates, as motivated by the COVID-19 pandemic." https://doi.org/10.1214/21-AOAS1499SUPP

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76**.

COCHRAN, J. J. (2020). Why we need more coronavirus tests than we think. *Significance* 14–15.

DE ANGELIS, D., PRESANIS, A. M., BIRRELL, P. J., TOMBA, G. S. and HOUSE, T. (2015). Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics* **10** 83–87.

DE VALPINE, P., TUREK, D., PACIOREK, C. J., ANDERSON-BERGMAN, C., TEMPLE LANG, D. and BODIK, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Statist.* **26** 403–413. MR3640196 https://doi.org/10.1080/10618600.2016.1172487

FAUST, J. S. (2020). Comparing COVID-19 deaths to flu deaths is like comparing apples to oranges. *Sci. Am.*.

FOG, A. (2008). Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. *Comm. Statist. Simulation Comput.* **37** 241–257. MR2422884 https://doi.org/10.1080/03610910701790236

GELMAN, A. and CARPENTER, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 1269–1283. MR4166866

GELMAN, A., RUBIN, D. B. et al. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GREENLAND, S. (2005). Multiple-bias modelling for analysis of observational data. *J. Roy. Statist. Soc. Ser. A* **168** 267–306. MR2119402 https://doi.org/10.1111/j.1467-985X.2004.00349.x

GREWELLE, R. and DE LEO, G. (2020). Estimating the global infection fatality rate of COVID-19. medRxiv.

GUDBJARTSSON, D. F., HELGASON, A., JONSSON, H., MAGNUSSON, O. T., MELSTED, P., NORDDAHL, G. L., SAEMUNDSDOTTIR, J., SIGURDSSON, A., SULEM, P. et al. (2020). Spread of SARS-CoV-2 in the Icelandic population. *N. Engl. J. Med.*

GUSTAFSON, P. (2010). Bayesian inference for partially identified models. *Int. J. Biostat.* **6** 17. MR2602560 https://doi.org/10.2202/1557-4679.1206

GUSTAFSON, P. and GREENLAND, S. (2009). Interval estimation for messy observational data. *Statist. Sci.* **24** 328–342. MR2757434 https://doi.org/10.1214/09-STS305

HALE, T., WEBSTER, S., PETHERICK, A., PHILLIPS, T. and KIRA, B. (2020). Oxford COVID-19 government response tracker. Available at https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker#data.

HAUSER, A., COUNOTTE, M. J., MARGOSSIAN, C. C., KONSTANTINOUDIS, G., LOW, N., ALTHAUS, C. L. and RIOU, J. (2020). Estimation of SARS-CoV-2 mortality during the early stages of an epidemic: A modelling study in Hubei, China and northern Italy. *PLoS Med.* **17** e1003189.

IOANNIDIS, J. (2020a). The infection fatality rate of COVID-19 inferred from seroprevalence data. (version 2 (June 8, 2020–14:00)). medRxiv.

IOANNIDIS, J. P. (2020b). First Opinion: A fiasco in the making? as the coronavirus pandemic takes hold, we are making decisions without reliable data. STAT. 2020. Available at https://tinyurl.com/uj539o4.

KOBAYASHI, T., JUNG, S.-M., LINTON, N. M., KINOSHITA, R., HAYASHI, K., MIYAMA, T., ANZAI, A., YANG, Y., YUAN, B. et al. (2020). Communicating the risk of death from novel coronavirus disease (COVID-19). *J. Clin. Med.* **9**. https://doi.org/10.3390/jcm9020580

KRUSCHKE, J. (2014). *Doing Bayesian Data Analysis*: *A Tutorial with R*, *JAGS*, *and Stan*. Academic Press, San Diego.

KÜMMERER, M., BERENS, P. and MACKE, J. (2020). A simple Bayesian analysis of the infection fatality rate in Gangelt, and an uncertainty aware extrapolation to infection-counts in Germany. Available at https://matthias-k.github.io/BayesianHeinsberg.html.

LAMBERT, P. C., SUTTON, A. J., BURTON, P. R., ABRAMS, K. R. and JONES, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat. Med.* **24** 2401–2428. MR2151713 https://doi.org/10.1002/sim.2112

LEE, G. (2020). Coronavirus: Why so many people are dying in Belgium. BBC.com. Available at https://www.bbc.com/news/world-europe-52491210.

LEON, D. A., SHKOLNIKOV, V. M., SMEETH, L., MAGNUS, P., PECHHOLDOVÁ, M. and JARVIS, C. I. (2020). COVID-19: A need for real-time monitoring of weekly excess deaths. *Lancet* **395** e81.

LI, S. and HUA, X. (2020). The closer to the Europe Union headquarters, the higher risk of COVID-19? Cautions regarding ecological studies of COVID-19. medRxiv.

LINTON, N. M., KOBAYASHI, T., YANG, Y., HAYASHI, K., AKHMETZHANOV, A. R., JUNG, S.-M., YUAN, B., KINOSHITA, R. and NISHIURA, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *J. Clin. Med.* **9** 538.

LIPSITCH, M. (2020). First Opinion: We know enough now to act decisively against COVID-19. social distancing is a good place to start. STAT. Available at https://tinyurl.com/yx4gf9mr.

LYONS, N. (1980). Closed expressions for noncentral hypergeometric probabilities. *Comm. Statist. Simulation Comput.* **9** 313–314.

MANSKI, C. F. (2003). *Partial Identification of Probability Distributions*. *Springer Series in Statistics*. Springer, New York. MR2151380

MEYEROWITZ-KATZ, G. and MERONE, L. (2020). A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates (version 4). medRxiv.

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. MR3937487 https://doi.org/10.1002/sim.8086

NEIL, M., FENTON, N., OSMAN, M. and MCLACHLAN, S. (2020). Bayesian network analysis of COVID-19 data reveals higher infection prevalence rates and lower fatality rates than widely reported. medRxiv.

O'DRISCOLL, M., DOS SANTOS, G. R., WANG, L., CUMMINGS, D. A., AZMAN, A. S., PAIREAU, J., FONTANET, A., CAUCHEMEZ, S. and SALJE, H. (2020). Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* 1–6.

ONDER, G., REZZA, G. and BRUSAFERRO, S. (2020). Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA* **323** 1775–1776. https://doi.org/10.1001/jama.2020.4683

OWID (2020). Codebook for the complete our world in data COVID-19 dataset. Available at https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data-codebook.md.

PASTOR-BARRIUSO, R., PÉREZ-GÓMEZ, B., HERNÁN, M. A., PÉREZ-OLMEDA, M., YOTTI, R., OTEO-IGLESIAS, J., SANMARTÍN, J. L., LEÓN-GÓMEZ, I., FERNÁNDEZ-GARCÍA, A. et al. (2020). Infection fatality risk for SARS-CoV-2 in community dwelling population of Spain: Nationwide seroepidemiological study. *BMJ* **371**.

PEARCE, N. (2000). The ecological fallacy strikes back. *J. Epidemiol. Community Health* **54** 326–327.

PELLIS, L., CAUCHEMEZ, S., FERGUSON, N. M. and FRASER, C. (2020). Systematic selection between age and household structure for models aimed at emerging epidemic predictions. *Nat. Commun.* **11** 1–11.

POLLÁN, M., PÉREZ-GÓMEZ, B., PASTOR-BARRIUSO, R., OTEO, J., HERNÁN, M. A., PÉREZ-OLMEDA, M., SANMARTÍN, J. L., FERNÁNDEZ-GARCÍA, A., CRUZ, I. et al. (2020). Prevalence of SARS-CoV-2 in Spain (ENE-COVID): A nationwide, population-based seroepidemiological study. *Lancet*.

PRESANIS, A. M., DE ANGELIS, D., THE NEW YORK CITY SWINE FLU INVESTIGATION TEAM, HAGY, A., REED, C., RILEY, S., COOPER, B. S., FINELLI, L. et al. (2009). The severity of pandemic H1N1 influenza in the United States. *PLoS Med.* **6**.

PROCHASKA, C. and THEODORE, L. (2018). Discrete probability distributions. *Introd. Math. Methods Environ. Eng. Sci.* 287.

RINALDI, G. and PARADISI, M. (2020). An empirical estimate of the infection fatality rate of COVID-19 from the first Italian outbreak. medRxiv.

SAHAI, H. and KHURSHID, A. (1995). *Statistics in Epidemiology*: *Methods*, *Techniques and Applications*. CRC press, Boca Raton.

SMEDT, T. D., MERRALL, E., MACINA, D., PEREZ-VILAR, S., ANDREWS, N. and BOLLAERTS, K. (2018). Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness. *PLoS ONE* **13** e0199180. https://doi.org/10.1371/journal.pone.0199180

STEVENS, W. L. (1951). Mean and variance of an entry in a contingency table. *Biometrika* **38** 468–470. MR0047287 https://doi.org/10.1093/biomet/38.3-4.468

STREECK, H., SCHULTE, B., KUEMMERER, B., RICHTER, E., HÖLLER, T., FUHRMANN, C., BARTOK, E., DOLSCHEID, R., BERGER, M. et al. (2020). Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. *Nat. Commun.* **11** 1–12.

SUTTON, A. J., COOPER, N. J., JONES, D. R., LAMBERT, P. C., THOMPSON, J. R. and ABRAMS, K. R. (2007). Evidence-based sample size calculations based upon updated meta-analysis. *Stat. Med.* **26** 2479–2500. MR2364400 https://doi.org/10.1002/sim.2704

THOMPSON, S. G. and HIGGINS, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Stat. Med.* **21** 1559–1573.

WONG, J. Y., HEATH KELLY, D. K., WU, J. T., LEUNG, G. M. and COWLING, B. J. (2013). Case fatality risk of influenza A (H1N1pdm09): A systematic review. *Epidemiology* **24**.

WU, J. T., LEUNG, K., BUSHMAN, M., KISHORE, N., NIEHUS, R., DE SALAZAR, P. M., COWLING, B. J., LIPSITCH, M. and LEUNG, G. M. (2020). Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* **26** 506–510. https://doi.org/10.1038/s41591-020-0822-7