Statistical estimation and decision-making for the COVID-19 pandemic

Nicholas J. Irons

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Adrian E. Raftery, Chair

Carlos Cinelli, Chair

Abel Rodriguez

Darryl Holman

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Statistical estimation and decision-making for the COVID-19 pandemic

Nicholas J. Irons

Co-Chairs of the Supervisory Committee:
Professor Adrian E. Raftery
Statistics and Sociology

Assistant Professor Carlos Cinelli
Statistics

This dissertation aims to provide policymakers and health practitioners with statistical tools and actionable information by which to make informed decisions, with a particular focus on the response to infectious disease outbreaks.

In the first project, we quantify how many Americans contracted COVID-19 in the first year of the pandemic. We formulate a Bayesian epidemiological model utilizing multiple sources of information, including random sample testing surveys, to debias clinical COVID data and estimate SARS-CoV-2 prevalence and transmission rates in the United States through March 2021. We quantify the extent to which reported COVID cases underestimated true infection counts, which was large (with about 2 in 3 infections missed by testing), especially in the first months.

Building on this work, in the second project we determine how to optimally respond to pandemics using non-pharmaceutical interventions (NPIs), which include social distancing measures, school and workplace closure, and testing, tracing, and masking policies. We first estimate the effects of NPIs on SARS-CoV-2 transmission in the US. Coupling these results with estimates of the costs associated to infections and NPIs derived from the public health

and economics literature, we evaluate the cost-effectiveness of NPI policies in the year prior to the arrival of COVID vaccines and antiviral treatments. Going further, we frame the problem of policy design in terms of statistical decision theory, with which we derive optimal NPI strategies. We find that pandemic school closures were not cost-effective, but other measures were.

In the third project, we propose a new method for the comparison of proportions—a foundational and ubiquitous statistical inference task relevant, in particular, to the analysis of randomized controlled trials with a binary outcome. Framing the problem as one of causal inference, we demonstrate how the likelihood can be cast in terms of clinically meaningful quantities, which facilitates interpretation, sensitivity analysis, and prior specification, and addresses the deficits of existing approaches. We demonstrate the utility of our method in empirical examples including a reanalysis of the Pfizer-BioNTech COVID-19 vaccine trial, which proved safe and highly efficacious in preventing SARS-CoV-2 infection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

BF: Bayes factor.

BHM: Bayesian hierarchical model.

BREASE: Baseline risk, efficacy, and adverse side effects.

CAR: Case ascertainment rate.

CDC: Centers for Disease Control and Prevention.

CDF: Cumulative distribution function.

COVID-19: Coronavirus disease 2019.

DA: Data augmentation.

DAG: Directed acyclic graph.

DALY: Disability-adjusted life-year.

GDP: Gross domestic product.

HMC: Hamiltonian Monte Carlo.

IB: Independent beta.

ICER: Incremental cost-effectiveness ratio.

IFR: Infection fatality rate.

JAGS: Just another Gibbs sampler.

JHU CSSE: Johns Hopkins University enter for Systems Science and Engineering.

LOESS: Locally estimated scatterplot smoothing.

LOO-CV: Leave-one-out cross validation.

LT: Logit transformation.

MAP: *Maximum a posteriori*.

MCMC: Markov chain Monte Carlo.

MRNA: Messenger ribonucleic acid.

NEJM: New England Journal of Medicine.

NPI: Non-pharmaceutical intervention.

NUTS: No-U-Turn sampler.

OC: Optimal control.

OXCGRT: Oxford COVID-19 government response tracker.

PCR: Polymerase chain reaction.

PDF: Probability density function.

PHS: Physician's Health Study.

PMF: Probability mass function.

QALY: Quality-adjusted life-year.

RT: Response type.

RT-PCR: Reverse transcription polymerase chain reaction.

SARS-COV-2: Severe acute respiratory syndrome coronavirus 2.

SEIRD: Susceptible-Exposed-Infectious-Removed-Deceased.

SIR: Susceptible-Infectious-Removed.

THAMES: Truncated harmonic mean estimator.

UNICEF: United Nations Children's Fund.

VSCD: Value of a statistical COVID-19 death.

VSL: Value of a statistical life.

WHO: World Health Organization.

ACKNOWLEDGMENTS

I would like to thank my research advisors and committee co-chairs, Adrian Raftery and Carlos Cinelli, for their patience, trust, encouragement, and tutelage. I have more praise to give these two brilliant and stately statisticians than I can fit here. I regularly pat myself on the back for having the good sense to seek them out.

Thank you to the members of my committee, Abel Rodriguez and Darryl Holman, who have graced me with their time and assistance and made this process as smooth as it could have been. And thank you, Abel, for your exemplary leadership of the department.

When I first sat down to write these acknowledgments, I made a list of all the people to whom I feel indebted for their support over the years. It was a good exercise in gratitude, but I arrived at the conclusion that it would be better to paint with broad brush strokes. So, for the remainder, I give thanks to groups. If you are reading this (highly unlikely), you know who you are.

Thank you to my community here at UW—the faculty, staff, and students of the Statistics and Biostatistics Departments, the Center for Statistics in the Social Sciences (CSSS), and the Center for Studies in Demography and Ecology (CSDE)—the sometimes invisible hands gently and consistently nudging me along.

Thank you to the members of the UW Statistics Working Group on Applied, Bayesian, and Computational Statistics for your years of valuable feedback, moral support, exceptional research, and engaging discussion.

Thank you to the funding sources that have paid my bills. My research was sup-

ported by NICHD grant number R01 HD070936, a Shanahan Endowment Fellowship, a Eunice Kennedy Shriver NICHD training grant, T32 HD101442-01, to CSDE, and the Boeing International Professorship at UW. I am especially grateful to the other organizers and participants in the NIH Training in Advanced Data Analytics for Behavioral and Social Sciences Research (TADA-BSSR) program.

Thank you to the Population Association of America for hosting what has become my favorite conference by far. I walk away from our meetings feeling energized, inspired, and excited for the future of population studies.

Thank you to my friends from the Max Planck Institute for Demographic Research in Rostock and Laboratoire MAP5 in Paris. You made me feel at home in Europe. And thank you to the faculty of the Oxford Department of Statistics and Leverhulme Center for Demographic Science for having faith in me.

Thank you to the passionate educators of Jean Baptiste Beaubien Elementary School, Northside College Preparatory High School, Northwestern University, and the University of Cambridge who have steered me in the course of my academic journey.

Thank you to my roommates, i.e., my Pacific Northwest family. Francis forever.

Thank you to my dear friends from Chicago, Evanston, Cambridge, Seattle, and elsewhere. And thank you to my family.

My last five years were a time of intense personal growth and becoming. I have learned so much from all of you. It takes a village. Thank you.

DEDICATION

to my loved ones, near and far

Chapter 1
INTRODUCTION

The global COVID-19 pandemic—the worst public health crisis in a century—tested our civilization's capacity to assess and respond to existential threats in real time. This dissertation, commenced and carried out during the pandemic, concerns fundamental questions pertinent to the monitoring and management of infectious disease transmission. Specifically, how do we quantify the gross health, social, and economic impacts of epidemics and the associated policy response? How do we assess the efficacy of public health interventions, pharmaceutical or otherwise? And how can we cost-effectively implement these interventions to optimally mitigate outbreaks? The remainder of this chapter summarizes our attempts to address these questions in a statistically principled manner and highlights our key contributions.

1.1 Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys

There are multiple sources of data giving information about the number of SARS-CoV-2 infections in the population, but all have major drawbacks, including biases and delayed reporting (National Academies of Sciences, Engineering, and Medicine, 2020). Representative random prevalence surveys, the only putatively unbiased source, are sparse in time and space, and the results can come with big delays. We develop a Bayesian framework to estimate viral prevalence by combining several of the main available data sources. It is based on a discrete-time Susceptible–Infected–Removed (SIR) model with time-varying reproductive

parameter. Our model includes likelihood components that incorporate data on deaths due to the virus, confirmed cases, and the number of tests administered on each day. We anchor our inference with data from random-sample testing surveys in Indiana and Ohio. We use the results from these two states to calibrate a model on positive test counts and proceed to estimate the infection fatality rate and the number of new infections on each day in each state in the United States between March 2020 and March 2021. We estimate the extent to which reported COVID cases have underestimated true infection counts, which was large, especially in the first months of the pandemic.

1.2 Evaluating and optimizing non-pharmaceutical interventions to combat infectious disease

Building on this work, we consider decision-making vis-á-vis non-pharmaceutical interventions. NPIs have proven to be effective tools for mitigating the spread of COVID-19 over the course of the pandemic. However, the individual effects of NPIs on viral transmission remain uncertain, which complicates decision-making concerning which policies to implement and when to loosen or tighten restrictions. Indeed, initial attempts to quantify the effects of NPIs were hampered by high correlation in the implementation of interventions early in the pandemic. With more data, subsequent studies have been able to disentangle the effects of NPIs at various geographic scales, although with substantial uncertainty remaining. Furthermore, the economic and social costs incurred by these interventions can be significant, including disrupted economic output, job losses, and learning loss.

We take a statistical data-driven approach to decision-making during the pandemic that weighs the effects of NPIs against their costs, in combination with a mechanistic epidemiological model, to navigate the economic, social, and public health trade-offs of COVID restrictions. We develop a Bayesian hierarchical epidemiological model to estimate the impacts of NPIs on SARS-CoV-2 transmission in the United States. Our model combines state-level data on cases and deaths due to COVID and intervention policies implemented

over time in all 50 states and D.C. We link these data—which track SARS-CoV-2 transmission and government intervention policies, respectively—via a regression function nested within a mechanistic epidemiological model simulating viral spread. We take a Bayesian approach to account for uncertainty in epidemiological parameters. Our model is able to capture the complex temporal dynamics of SARS-CoV-2 transmission, providing validation for its use in NPI policy planning. Going further, we combine this model with estimates of the costs of interventions from the literature to quantitatively evaluate those policies that have been implemented during the pandemic and to formulate more effective and less costly NPI strategies.

1.3 Causally sound priors for binary experiments

In this chapter, we propose a new prior for the Bayesian comparison of proportions, which is a common task in the analysis of randomized controlled trials with a binary outcome. Because the data are derived from a randomized experiment, the main existing Bayesian methods ignore the causal structure underlying the data generating process. Approaching the problem from a causal inference perspective, we propose a novel parametrization of the likelihood in terms of counterfactual quantities derived from the potential outcomes—specifically, the baseline risk (in the control group), and the efficacy and risk of adverse side effects of treatment (BREASE). These parameters, which are familiar from the clinical vocabulary, facilitate interpretation, prior elicitation, and sensitivity analysis of treatment effects.

As the three parameters in our model are variation independent proportions, we propose a joint independent beta prior, which we show to have a number of advantages relative to existing methods. After introducing the method, we re-analyze the results of the aspirin component of the 1980s Physicians' Health Study (PHS), a landmark trial that contributed to the widespread prescription of low-dose aspirin for prevention of myocardial infarction. Numerous subsequent trials have found that aspirin has a small effect, if any, for the primary

prevention of cardiovascular events, but that it significantly increases the risk of major hemorrhage, particularly in the older age groups most commonly prescribed. Frequentist and Bayesian methods produce markedly different conclusions from the PHS: p-values indicate strong evidence for an effect of treatment, while Bayes factors for the two most common priors yield either strong evidence in favor of the null (no effect), or moderate evidence for the alternative. Using the BREASE approach, we demonstrate how to reconcile these disparate results, clarify what one needs to believe in order to claim that a treatment is effective, and transparently distinguish robust from fragile findings. We find that the results of the trial are ambiguous, and the conclusion that aspirin is effective for primary prevention of fatal heart attack strongly depends on the prior. This need not always be the case: in a reanalysis of the Pfizer-BioNTech COVID-19 vaccine trial, we find robust, unequivocal evidence that the vaccine is highly efficacious.

1.4 Outline of the dissertation and contributions

Chapter 2 details our model to estimate SARS-CoV-2 infections in the US during the first year of the pandemic from deaths, confirmed cases, tests, and random surveys. We find that the large majority of SARS-CoV-2 infections went undetected, with case ascertainment exceptionally low in the spring of 2020 and improving over time. Furthermore, we provide the first (and, in some cases, only) estimates of the COVID-19 infection fatality rate (IFR)—a key parameter in infectious disease modeling—in nearly every US state. Our paper, published in *Proceedings of the National Academy of Sciences*, has been highly cited, received extensive media coverage, and has been used by government officials. As we discuss in Chapters 2 and 3, our estimates, which were produced early in the pandemic using limited data, have stood the test of time and are in line with the findings of subsequent COVID prevalence studies in the US and elsewhere based on more extensive data, including other random testing surveys. Knowledge of SARS-CoV-2 incidence is essential for accurately tracking

viral spread, informing public policy, assessing the effectiveness of public health measures, and determining the burden of disease in a population, among other reasons. As such, our estimates serve as a necessary input to our study of non-pharmaceutical interventions in Chapter 3.

Chapter 3 extends the model developed in Chapter 2 in order to evaluate and optimize NPIs to combat infectious disease. Our methodology builds upon and extends the literature studying NPI mitigation of infectious disease in order to address a number of gaps that severely limit its value in informing and evaluating policies. Firstly, to the best of our knowledge, we are the first to phrase the control of infectious disease transmission as a statistical decision problem. In particular, this means that our conclusions are drawn from observed data rather than the output of simulation models and that we account for parameter uncertainty in a statistically coherent way. Secondly, we model the costs and effects on viral transmission of multiple specific NPIs calibrated to U.S. economic and state-level COVID data. Modeling papers studying the cost-effectiveness of NPIs and their optimal control of pandemics often consider a limited toolkit. They tend to focus on a minimal collection of interventions, often a single NPI, which, in the economics literature, usually represents a catch-all "social distancing", "containment", or "lock-down" policy. This is impractical, as we generally have a range of tools at our disposal, and NPIs are known to be more effective—and, therefore, more cost-effective—in combination. Furthermore, if we consider only a single instrument, we may erroneously conclude that its implementation is cost-effective because we implicitly assume that other policies either cannot be enacted or cannot be modified from a set schedule. As such, our findings are more precise and informative than the broad qualitative guidance drawn from prior studies in the context of COVID-19—e.g., that "lock-down" is cost-effective and optimal when implemented early and stringently. Thirdly, we depart from the large majority of studies assessing the economic impact and cost-effectiveness of school closure during pandemics by factoring in costs associated

to student learning loss. Additional methodological contributions of our approach include a novel zero-inflated negative binomial model that flexibly captures well-known reporting idiosyncrasies and over-dispersion in clinical COVID data. As a result, our method eliminates the need for ad hoc data cleaning and smoothing procedures that can complicate the analysis pipeline, yield poorly calibrated prediction intervals, and potentially bias transmission rate estimates based on over-smoothed data. Furthermore, we implement a two-stage modeling procedure that first estimates the time-varying effective reproduction number in each US state individually, followed by a joint hierarchical model across states that estimates pooled effects of NPIs on transmission dynamics. This approach allows for efficient Bayesian computation by parallelizing model fits across states. Finally, our methodology has implications for the use of incremental cost-effectiveness ratios (ICERs) in infectious disease.

We find that, although school closures significantly reduced viral transmission, their social impact in terms of student learning loss was far too costly. Conditional on the other policies enacted, extended school closures imposed a cost to the nation's youth in service of its older generations, reducing the latter's risk of death at the expense of $2 trillion (USD2020) in future GDP. Moreover, we find that this marginal trade-off between school closure and COVID deaths was not inescapable: more timely, stringent, and enduring use of other measures in combination would have sufficed to maintain similar or lower mortality rates without incurring profound learning loss. Optimal NPI strategies involve consistent implementation of mask mandates, public test availability, contact tracing, social distancing orders, and reactive workplace closures, with no closure of schools beyond the usual 16 weeks of break per year. Their use would have reduced the gross impact of the pandemic in the U.S. in 2020 from $5.1 trillion to $2.4 trillion and, with high probability, saved lives.

While our study focuses on COVID-19 in the U.S. prior to the arrival of vaccines, our qualitative findings shed light on NPI implementation in other settings. Masking, testing, and tracing are relatively cheap and likely to remain cost-effective universally: for severe

and relatively mild pandemics; in lower resource settings; and after effective pharmaceutical interventions become available. After the arrival of vaccines and antiviral treatments, workplace closures and social distancing measures should be enacted more sparingly. Although school closures were not cost-effective, evidence suggests that distance learning helped to mitigate learning loss. Consequently, extended school closures are likely to be relatively more costly in low- and middle-income countries with younger populations and less capacity to provide effective education remotely. Likewise, with fewer opportunities for remote work and less online economic activity, workplace closures, stay-at-home orders, and other social distancing measures may be more costly in these countries. For less virulent diseases with a similar age pattern of death, extended school closures should never be implemented and extended workplace closures and social distancing measures should be mandated with care. If possible, more targeted interventions should be utilized.

Chapter 4 introduces the BREASE model, our novel method for the Bayesian comparison of proportions, a common and centuries-old statistical inference problem. BREASE has a number of desirable features relative to the two main existing approaches. Our proposal: induces prior dependence between expected outcomes in the treatment and control groups in a principled manner; frames the model in terms of quantities inherently familiar to clinicians, which facilitates interpretability of the parameters, elicitation of prior knowledge, and sensitivity analysis; is analytically tractable, with formulae for the marginal likelihood, Bayes factor, and other posterior quantities, as well as exact posterior sampling via simulation, in cases where traditional MCMC fails; and results in a prior under the alternative that concentrates mass on the nested null hypothesis of no average treatment effect, which yields favorable performance for Bayesian hypothesis testing and does not succumb to the Jeffreys-Lindley paradox.

Chapter 5 presents concluding remarks and future directions.

Chapter 2

ESTIMATING SARS-COV-2 INFECTIONS FROM DEATHS, CONFIRMED CASES, TESTS, AND RANDOM SURVEYS

As of the writing of this chapter in July 2021, the novel coronavirus SARS-CoV-2 had infected over 33 million people in the United States. Nationwide over 600,000 had died in the COVID-19 pandemic, which has necessitated shutdowns of schools and sectors of the economy. Nevertheless, the extent of the virus' spread remains uncertain. There are multiple sources of data giving information about the number of SARS-CoV-2 infections in the population, but all have major drawbacks, including biases and delayed reporting. For example, the number of confirmed cases largely underestimates the number of infections, deaths lag infections substantially, while test positivity rates tend to greatly overestimate prevalence. Representative random prevalence surveys, the only putatively unbiased source, are sparse in time and space, and the results can come with big delays. Reliable estimates of population prevalence are necessary for understanding the spread of the virus and the effectiveness of mitigation strategies. We develop a simple Bayesian framework to estimate viral prevalence by combining several of the main available data sources. It is based on a discrete-time SIR model with time-varying reproductive parameter. Our model includes likelihood components that incorporate data on deaths due to the virus, confirmed cases, and the number of tests administered on each day. We anchor our inference with data from random sample testing surveys in Indiana and Ohio. We use the results from these two states to calibrate the model on positive test counts and proceed to estimate the infection fatality rate and the number of new infections on each day in each state in the USA. We estimate the extent to which reported COVID cases have underestimated true infection counts, which was

large, especially in the first months of the pandemic. Specifically, we find that approximately 60% of infections have gone unreported in the first year of the pandemic. Even so, only about 20% of the US had been infected as of early March 2021, suggesting that the country was far from herd immunity at that point.

This chapter is based on the article "Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys" published in *Proceedings of the National Academy of Sciences* (Irons & Raftery, 2021).

2.1 Introduction

SARS-CoV-2 test data are fraught with biases that obscure the true rate of infection in the population. Lack of access to viral tests, which was particularly pronounced in the early days of the pandemic, in conjunction with selection bias due to asymptomatic and mild infections, yield case counts that tend to underestimate the true number of infections in the population. By the same token, test positivity rates tend to overestimate viral prevalence. Hospitalization rates and emergency room visits do not estimate the overall infection rate, and are not comparable between states or counties, or over time. Reported deaths due to COVID are considered less problematic as an estimate of the true death count and provide a more accurate reflection of the course of the pandemic (National Academies of Sciences, Engineering, and Medicine, 2020).

We combine several of the main sources of data relevant to the number of infections using a simple Bayesian model that accounts for the biases and delays in the data. Our model relies on data on deaths due to COVID, confirmed cases, and testing reported by the COVID Tracking Project (The Atlantic, 2021). We use a modified Susceptible-Infected-Removed (SIR) model, a compartmental epidemiological model widely used to simulate the spread of disease in a population (Kermack & McKendrick, 1927). We combine this with a Poisson likelihood for death counts and a normal likelihood for estimates of viral and

seroprevalence from random sample testing surveys conducted in Indiana and Ohio (Kline et al., 2021; Menachemi et al., 2020).

With these data we infer the infection fatality rate (IFR) and obtain statistically principled estimates of the number of new infections on each day since March 2020 in Indiana and Ohio. We then leverage our results from these states to build a model for confirmed cases that accounts for preferential testing as a function of the cumulative number of tests administered in each state. This allows us to pin down the IFR and infection counts for the vast majority of states that have not conducted representative testing surveys.

Our simple Bayesian model takes inspiration from Johndrow et al. (2020), although it differs in significant ways. Whereas Johndrow et al. (2020) model the effect of social distancing measures by allowing the SIR contact parameter to change pre- and post-lockdown, we allow it to vary in time to account for fluctuation in the tightening and loosening of restrictions, as well as in adherence to the restrictions. Furthermore, we incorporate testing data, develop a statistical model for preferential testing, and include the IFR as a parameter in the model to be estimated, rather than a fixed constant. Finally, to simplify model implementation we use a discrete time SIR model, rather than a continuous time model based on differential equations.

2.2 Methods

2.2.1 SIR model

We first define our discrete-time SIR model for infections in each state. Let S_t denote the number of susceptible people in the population on day t, I_t the number of infections, and R_t the number removed. The number removed includes those who have died of the disease and those who have recovered, and are assumed immune for the rest of the period of our study. With N denoting the state population, these quantities evolve in time according to

the equations

$$\begin{cases} S_{t+1} - S_t &= -\frac{\beta_t}{N} I_t S_t, \\ I_{t+1} - I_t &= \frac{\beta_t}{N} I_t S_t - \gamma I_t, \\ R_{t+1} - R_t &= \gamma I_t. \end{cases} \tag{2.1}$$

Note that $\nu_t = S_{t-1} - S_t$ is the number of new infections on day t. We allow the parameters β_t, interpreted as the mean number of contacts per person on day t, to vary over time. This accounts for variation in exposure due to implementation or loosening of social distancing and other policy measures over time. We model β_t as a random walk with step size σ estimated from the data, $\beta_{t+1} \sim \text{Normal}(\beta_t, \sigma^2)$. We assume that γ^{-1}, the average length in days of the infectious period, is determined by the disease and is therefore constant over time.

2.2.2 Likelihood on deaths

Let $\tau = \{\tau_0, \tau_1, \ldots, \tau_m\}$ denote the distribution of time to death for those infected individuals who die from the disease, i.e., τ_s is the probability of death s days after infection, conditional on death occurring. Similar to Johndrow et al. (2020), who calibrated τ by matching quantiles of a negative binomial distribution to case data from China (Lauer et al., 2020a; Zhou et al., 2020), we assume that τ follows a NegativeBinomial$(\alpha, 1/(\beta + 1))$ distribution with parameters $\alpha = 21, \beta = 1.1$, and we truncate the distribution at the 99th percentile, or $m = 40$ days, to rule out extremely delayed deaths. We denote by D_t the reported deaths due to COVID on day t, which we obtain from the COVID Tracking Project (The Atlantic, 2021). We link the daily new infection counts $\nu = (\nu_t)_t$ to reported deaths via the likelihood $D_t \overset{\text{ind.}}{\sim} \text{Poisson}\left(\text{IFR} \sum_{k=1}^{t} \nu_k \tau_{t-k}\right)$.

2.2.3 Representative random prevalence surveys

To pin down the IFR, we add likelihood components incorporating the Indiana and Ohio prevalence survey data (Kline et al., 2021; Menachemi et al., 2020). Active viral prevalence in Indiana in the period April 25–29, 2020 was estimated as $\hat{\theta}_v = 1.74\%$. We model this quantity using a normal approximation to the binomial distribution, $\hat{\theta}_v \sim \text{Normal}\left(\theta_v, \frac{\theta_v(1-\theta_v)}{n_v}\right)$, where $\theta_v = (\sum_{t=T_1}^{T_2} I_t)/N(T_1 - T_2)$ is the average viral prevalence between days $T_1 =$ April 25 and $T_2 =$ April 29. Here $n_v = 3,605$ is the number of viral tests administered. Similarly, the estimated seroprevalence in the testing period, $\hat{\theta}_s = 1.09\%$, is modeled as $\hat{\theta}_s \sim \text{Normal}\left(\theta_s, \frac{\theta_s(1-\theta_s)}{n_s}\right)$, where $\theta_s = \sum_{t=T_1}^{T_2} R_t/N(T_1 - T_2)$ and $n_s = 3518$. These results come from the first phase of the Indiana prevalence survey described in Menachemi et al. (2020). The sampled population consisted of all non-institutionalized Indiana residents aged ≥ 12 years listed on state tax returns, including filers and dependents. Stratified random sampling was conducted using Indiana's 10 public health preparedness districts as sampling strata, and 15,495 participants were contacted by the state health department. Of those contacted, 3,658 or 23.6% agreed to participate in the study. While low, this response rate is not far from the survey industry average of 30% (Dixon & Tucker, 2010; Holbrook et al., 2007). Menachemi et al. (2020) note that respondents might have been subject to response bias, which could have resulted in underestimates or overestimates. To adjust for differences in nonresponse between groups, data were weighted for age, race, and Hispanic ethnicity. Participants were tested for active infection via RT-PCR and past infection via antibody test between April 25 and April 29, 2020. The RT-PCR tests used had high but imperfect sensitivity, and the antibody tests had high but imperfect specificity. The former could have caused false negative results and the latter false positive results. In a follow-up paper published in PNAS, Yiannoutsos et al. (2021) conducted a Bayesian analysis of the Indiana survey data to address uncertainty in the results related to imperfect testing and difference in prevalence among subgroups characterized by ethnicity, race, and age. Due to very low

response rates—less than 8% in the second and third phases—we do not include data from the subsequent phases of the Indiana study in our analysis.

The likelihood for the prevalence survey data from Ohio is analogous. The survey design was a stratified two-stage cluster sample, with strata defined by 8 administrative regions in the state of Ohio. Within each region, 30 census tracts were randomly selected with probability proportional to total population. Within each tract, households were randomly sampled and one adult within each household was randomly selected to participate in the study. Of those contacted, 727 or 18.5% agreed to participate. Between July 9 and July 28, 2020, participants were tested for active and past infection via RT-PCR and antibody test. Due to the low response rate and imperfect diagnostic tests used in the study, the same caveats described above for the Indiana survey apply. Kline et al. (2021) conducted a Bayesian analysis of the seroprevalence survey data to address the uncertainty in the results associated with nonresponse and imperfect testing. As reported in Kline et al. (2021), the estimated seroprevalence in the state was $\hat{\theta}_s = 1.3\%$ in the period July 9–28, with a sample size of $n_s = 667$. Results from the PCR tests in the same study were reported in a press conference on October 1 available on YouTube (The Ohio Channel, 2020). The viral prevalence in that period is estimated as $\hat{\theta}_v = 0.9\%$ with sample size $n_v = 727$. To the best of our knowledge, these numbers have not yet been published.

2.2.4 Modeling preferential testing

As shown in Figures 2.2 and 2.3, the undercount curve $(I_t + R_t)/(\sum_{k \le t} C_k)$ has a common shape in Indiana and Ohio. Here, I_t and R_t are the SIR parameters on day t, and C_t is the total number of confirmed and probable cases, defined as unique people with a positive PCR or other approved nucleic acid amplification test in the state on day t, as reported by the COVID Tracking Project (The Atlantic, 2021). We found that the reciprocal of the undercount is approximately linear when plotted against the square root of the cumulative

number of tests administered in the state on each day, and that the slopes of these lines for the two states are similar; see Figure 2.1. This led to the following model for the test data:

$$\sum_{k=1}^{t} C_k \sim \text{Normal}\left(\phi_t(I_t + R_t), \eta_t^2\right). \tag{2.2}$$

Here the parameters ϕ_t and η_t are proportional to the square root of the fraction of the population tested up to day t,

$$\phi_t = \phi\sqrt{\frac{\sum_{k=1}^{t} T_k}{N}}, \qquad \eta_t^2 = \eta^2 \frac{\sum_{k=1}^{t} T_k}{N},$$

so that ϕ_t is the overall fraction of infections that appear in the cumulative number of positive tests. We assume that this fraction grows as the state's test capacity ramps up and that the variance in this relationship, η_t^2, grows linearly with the total number of tests administered. Here T_t is the number of total test results in the state on day t as reported by the COVID Tracking Project (The Atlantic, 2021). Due to variation in test reporting methods across states, this number may include antigen tests as well as viral (PCR) tests. Moreover, different states report total tests using different units, whether in terms of test encounters, test specimens, or unique people tested. As such, T_t is best understood as an estimate of the state's test capacity. This is the extent to which it is used in our preferential testing model. For example, we do not model test positivity rates C_t/T_t on each day.

To arrive at the distribution in (2.2), we can model the cases on each day independently as

$$C_t \overset{ind.}{\sim} \text{Normal}\left(\phi_t(I_t + R_t) - \phi_{t-1}(I_{t-1} + R_{t-1}), \eta^2 \frac{T_t}{N}\right). \tag{2.3}$$

Noting that $\nu_t = (I_t + R_t) - (I_{t-1} + R_{t-1})$, we can write the mean of C_t as

$$\phi_t \cdot \nu_t + (\phi_t - \phi_{t-1})(I_{t-1} + R_{t-1}).$$

Hence, in expectation C_t can be decomposed as a fraction of the new infections on day t, ν_t, and a smaller fraction of the cumulative incidence on day $t-1$, $I_{t-1} + R_{t-1}$.

In fitting the model, we do not use the likelihood on each day (2.3) due to inconsistent reporting of cases and tests, as well as weekly oscillations in these numbers due to reduced reporting on weekends. Rather, in each state we combine cases and tests into non-overlapping consecutive L-day periods, where L is at least 7 to account for weekend effects, and model the counts in these periods independently.

We first fit the model in Indiana and Ohio without the likelihood on cases described above. That is, initially we used only deaths data and the random sample surveys in each state. With the resulting posterior samples of cumulative incidence $I_t + R_t$ on each day, we arrived at the likelihood on cases. Figure 2.1 demonstrates the relationships defined in equations (2.2) and (2.3). We refer to the normal means in (2.2) and (2.3) (divided by the parameter ϕ) as the cumulative and marginal regression functions, respectively. The lower panels of Figure 2.1 reveal a comparable slope ϕ for Indiana and Ohio after a brief initial period when testing and cases were very low. The widening confidence intervals in the upper panels exhibit the growth of the variance in (2.2) as a function of cumulative testing.

A number of other models for case and test data have been proposed. Campbell et al. (2022) introduced a binomial likelihood on cases, $C_t \sim \text{Binomial}(T_t, 1-(1-I_t/N)^\alpha)$, where I_t/N is the infection rate on day t and $\alpha > 0$ is a parameter representing the degree of preferential testing. Assuming the infection rate is small, a binomial expansion of the test positivity rate yields the approximation $1-(1-I_t/N)^\alpha \approx \alpha I_t/N$. An application of Bayes' rule to the latter model shows that $\alpha = P(\text{tested}|\text{infected})/P(\text{tested})$. This model has some limitations in the context of our study. Firstly, the degree of preferential testing α is likely to decrease as testing increases, and it is not obvious how one might parametrize $\alpha = \alpha_t$ to account for this. Secondly, the model is not additive, as the test positivity relies on the active infection rate. As a result, it is not well suited to handling state-level testing data, which can be unreliable on the daily level.

Gu (2021) and Ellis (2020) proposed similar models to correct case counts using test

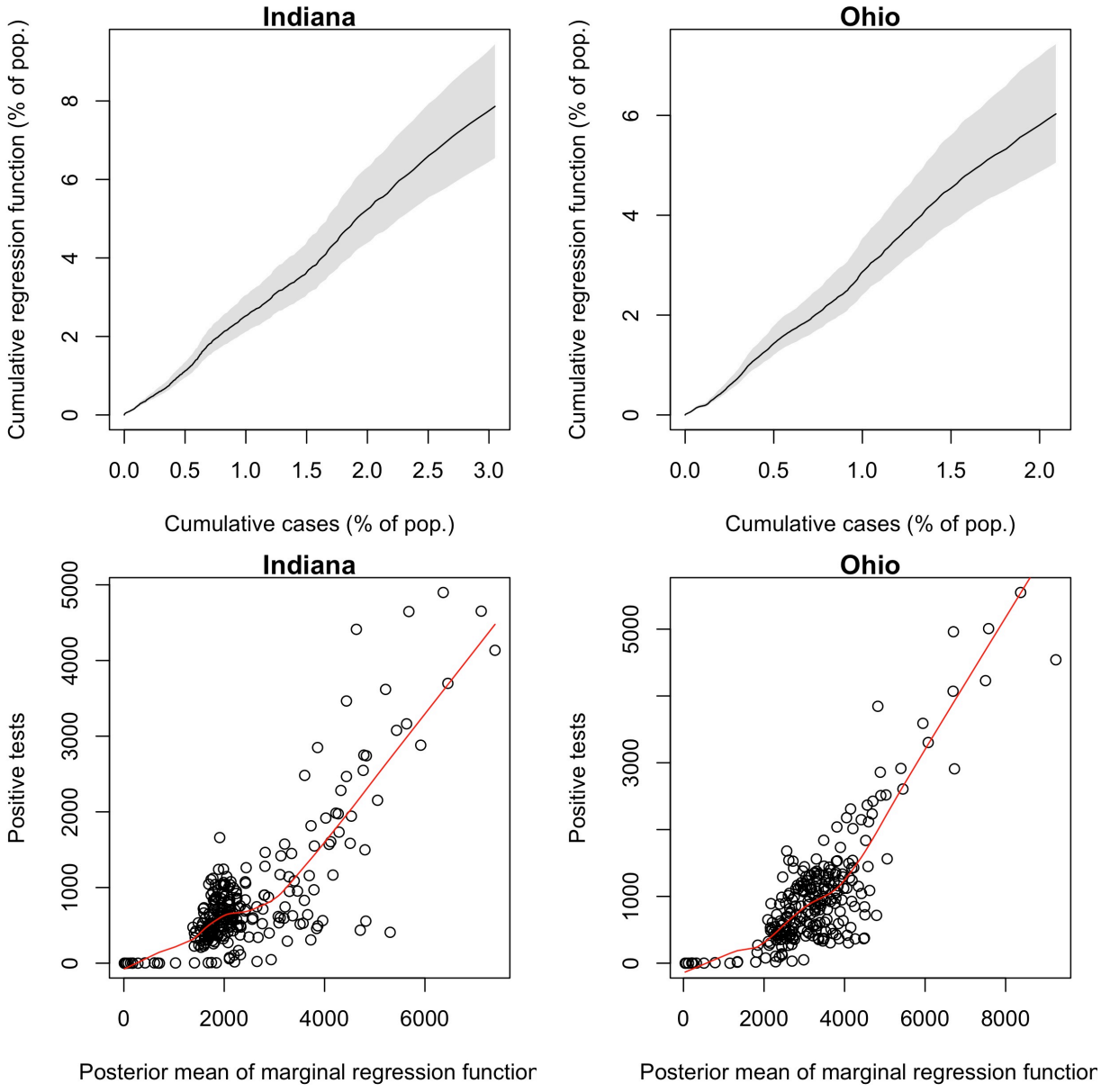


Figure 2.1: Upper panels: Posterior median and 95% confidence bands for the cumulative regression function in equation (2.2) plotted against cumulative cases in Indiana and Ohio. Lower panels: Positive tests on each day plotted against the posterior mean of the marginal regression function in equation (2.3). LOESS curves are plotted in red.

positivity rates. They take the form $\nu_t = C_t[m \cdot (C_t/T_t)^k + b]$ where $m > 0, k \in [0, 1], b \geq 0$ are parameters. Benatia et al. (2020) also estimate population prevalence on day t by the number of positive tests on day t scaled by a multiplicative factor depending on the number of tests administered on day t as a fraction of the state population. These models are susceptible to the same issues as that of Campbell et al. They rely on daily test positivity rates, which are reported inconsistently across states (Kissane & Rivera, 2020). And as Gu (2021) notes, the parameters estimated at one point in time do not carry over to other time periods. Furthermore, by assuming that new infections are a function only of cases and tests on that day, these models ignore the lag between infections and their confirmation via testing. They also presume that there are no new infections on days in which no positive tests are reported. Our likelihood on cases in equation (2.3) allows for new infections to be reflected in case counts at a later date.

Note that our model does not take into account imperfect testing. Modeling imperfect testing is complicated by the inconsistent test reporting methods across states described above, which obscures the true number of PCR tests administered in a state on each day. Given that estimated active infection rates are generally low ($< 5\%$) at any given time, imperfect test specificity (i.e., the proportion of true negative results) is a greater potential source of bias in case counts C_t than sensitivity. False positives resulting from imperfect specificity would increase C_t. We note, however, that the molecular RT-PCR assays widely deployed to test for the presence of viral RNA are shown to have near perfect specificity (Böger et al., 2021; He, Luo, et al., 2020; Xie et al., 2020; Yu et al., 2020).

2.2.5 Prior specification

Lastly, we specify prior distributions for the model parameters $\{\text{IFR}, \beta_1, \sigma, \gamma^{-1}, (S_1, I_1), \phi, \eta\}$. We used a weakly informative Uniform$(0, 0.03)$ prior distribution for the IFR in each state. For Indiana, we used a truncated normal prior for the mean infectious period, $\gamma^{-1} \sim$

$\text{Normal}_{[5.5,11.5]}(8.5, 1.5^2)$. This is motivated by clinical data, which show that most infected individuals remain infectious no longer than 10 days after symptom onset (Arons et al., 2020; Bullard et al., 2020; CDC, 2020; Korea Centers for Disease Control and Prevention, 2020; Lu et al., 2020; Quicke et al., 2020; van Kampen et al., 2020; Wölfel et al., 2020), and that patients can be highly infectious several days before symptom onset (He, Lau, et al., 2020).

We assumed that the removal rate γ is determined by the disease and so does not vary between states. Therefore, after fitting the model to the data for Indiana, we used the posterior distribution of γ for Indiana as the prior distribution of γ for Ohio. We then used the posterior distribution from Ohio as the prior distribution for the remaining states, each of which we modeled independently. To estimate the preferential testing parameter ϕ, we used the same process as described for γ. In Indiana, the prior is uninformative, $\phi \sim \text{Uniform}(0, 2)$. We use the following weakly or uninformative independent uniform priors for the remaining parameters:

$$\begin{aligned}
\text{IFR} &\sim \text{Uniform}(0, 0.03), \\
\beta_1 &\sim \text{Uniform}(0, 1), \\
\sigma &\sim \text{Uniform}(0, 1), \\
\eta &\sim \text{Uniform}(0, 3), \\
(S_1, I_1, R_1) &\sim 0.01N \cdot \text{Dirichlet}(1, 1, 1) + (0.99N, 0, 0).
\end{aligned}$$

In the last line N denotes the state population. This distribution specifies that no more than 1% of the state has been infected as of the first day for which COVID Tracking Project data are available for the state (i.e., the susceptible population is $S_1 \geq 0.99N$), which is generally between late February and mid-March 2020. Note that the $\text{Dirichlet}(1, 1, 1)$ distribution is the uniform distribution on the 2-dimensional simplex. While the contact parameter on the first day β_1 is given an uninformative uniform prior, on subsequent days we assume that β_t follows a random walk model, $\beta_{t+1} \sim \text{Normal}(\beta_t, \sigma^2)$.

2.2.6 Implementation

We built the model in R and fit it with the RStan software package, which implements the No-U-Turn Sampler for Bayesian inference (Hoffman, Gelman, et al., 2014; R Core Team, 2020; Stan Development Team, 2020). For each state, we ran 4 chains in parallel for 20,000 steps each with the first 10,000 as burn-in to obtain 40,000 samples from the posterior distribution of the model parameters. Code to fit the model is available at the GitHub repository (https://github.com/njirons/covidest).

2.2.7 Data cleaning

In certain states, the COVID Tracking Project data reports a negative number of cases, tests, or deaths on some days, often due to record de-duplication or changes in data reporting by the state government. If a negative number of cases or tests is reported, we address this by setting that datum to zero and distributing the negative number over all previous days proportional to the number of cases or tests reported on those days. If a negative number of deaths is reported, we set that datum to zero and subtract the negative number from the deaths reported on the previous day. If this results in a negative number of deaths on the previous day, we continue this procedure until all counts are non-negative.

The COVID Tracking Project also notes days when state governments report a backlog of cases or deaths, which usually results in a large spike in the data on that day. We address this by setting that datum to the average of the number of cases or deaths reported on the day before and the day after, and distributing the excess number of cases or deaths over all previous days proportional to the number of cases or deaths reported on those days.

2.3 Results

Here we present detailed results for Indiana and Ohio, two states with statewide representative random sample testing surveys that we incorporate into our probabilistic model. In

order to assess the accuracy of our estimates, we also present results for Connecticut and New York. Connecticut has conducted a statewide representative seroprevalence survey (Mahajan et al., 2021) and was also included in a nonrandom seroprevalence study of 10 sites across the country (Havers et al., 2020). New York has the highest number of reported deaths due to COVID and there is a body of literature studying the spread of the disease in the state, including New York City Health Department (2021), Stadlbauer et al. (2021), and Yang, Kandula, et al. (2021). The estimates from these studies provide a basis of comparison for our results.

We also present aggregated estimates for the entire United States. Tables A.1 and A.2 in the Appendix include estimates of the IFR and the cumulative incidence (i.e, the percent of the state's population having been infected) and undercount factor for all 50 states and the District of Columbia as of March 7, 2021, the last day reported by the COVID Tracking Project. Plots for the 50 states and DC are also shown in the Appendix.

2.3.1 Indiana

We estimate an IFR of 0.84% (95% interval 0.70–1.00) in Indiana. We estimate the cumulative incidence of COVID-19 in the state at 19.7% (16.5–23.7) as of January 1, 2021, and 20.9% (17.5–25.1) as of January 15, 2021. These numbers agree with internal estimates provided by Constantin Yiannoutsos and Justin Blackburn in correspondence, which are based on inverting the IFR reported by Blackburn et al. (2021). Their estimates are 18.9% by January 1, 2021 and 21.2% by January 15, 2021. By March 7, 2021, cumulative incidence had increased to 22.9% (19.2–27.6), nearly a quarter of the state, or about 1.5 million infections. There have been 2.3 (1.9–2.8) infections for every confirmed case in the state through this date. This suggests that a large majority of infections in the course of the pandemic have gone unreported, although Figure 1 shows that undercounting was most pronounced early on and has improved substantially over time.

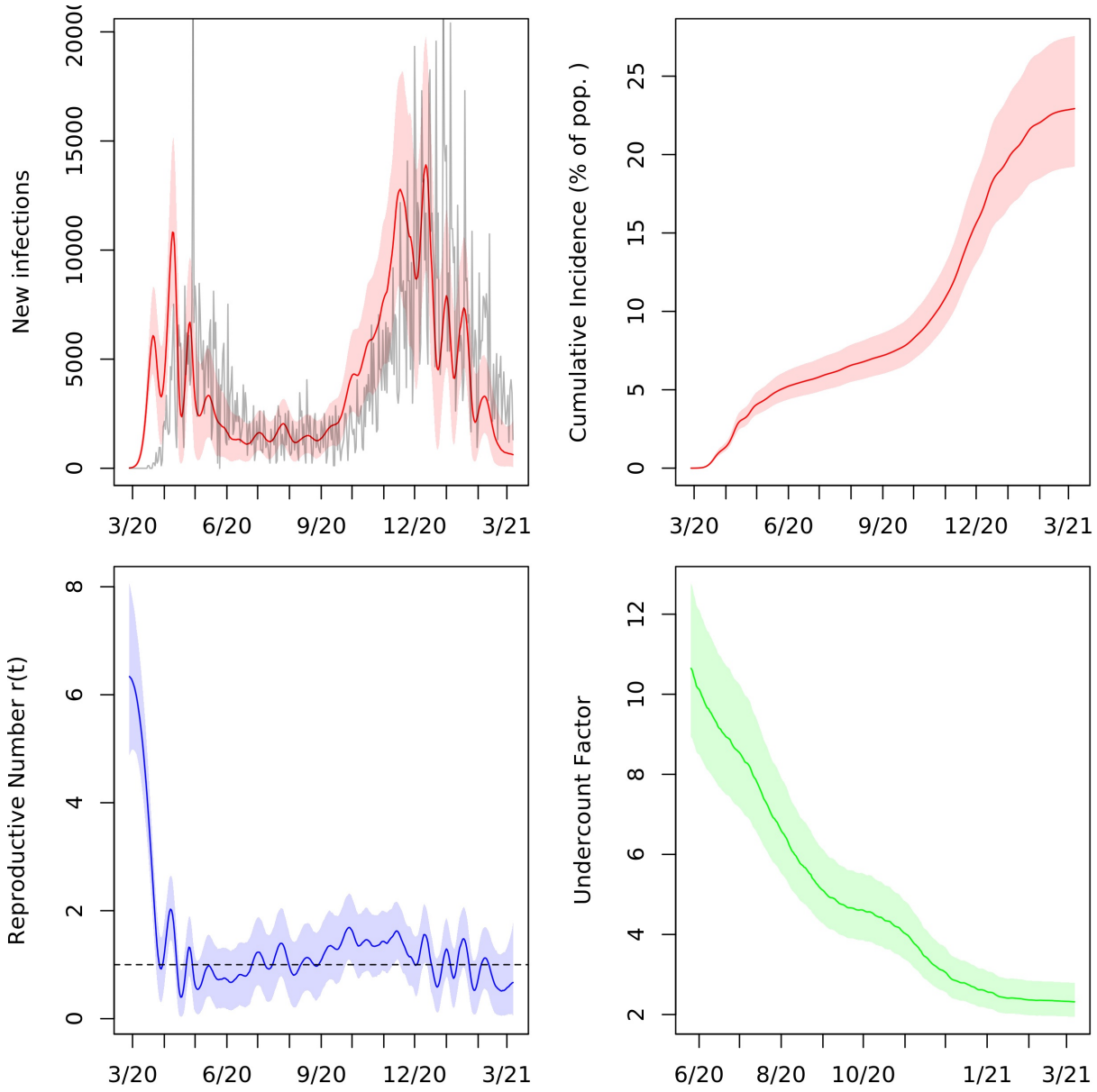


Figure 2.2: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, reproductive number $r(t)$, and cumulative undercount in Indiana from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

Figure 1 exhibits posterior estimates of new infections on each day, ν_t, as well as the cumulative undercount factor, which is the ratio of estimated cumulative infections to cumulative confirmed cases. Figure 2.2 displays the viral prevalence, the cumulative incidence, and the reproductive number $r(t) = \beta_t/\gamma$ on each day.

By the time that the first confirmed case was reported in Indiana on March 6, 2020, there had likely been more than 800 infections in the state (95% interval 483–1,384). We estimate that as of May 1, 2020, there were 274,000 cumulative infections (95% interval 230,000–327,000), compared to 18,630 confirmed cases by that date. This yields a cumulative incidence of 4.1% (3.4–4.8) and an undercount factor of 14.7 (12.4–17.6). This estimate is comparable to others in the literature for that period (Havers et al., 2020; Johndrow et al., 2020; Wu et al., 2020). Between the 16th and 19th of March 2020, the state's Governor Eric Holcomb ordered a stop to indoor dining, declared a state of emergency, and closed schools; on March 23rd 2020 he issued a stay-at-home order. According to our model, the first wave of infections reached its peak about two weeks later in early April 2020.

2.3.2 Ohio

We estimate an IFR of 0.83% (95% interval 0.68–1.03) in Ohio. As of March 7, 2021, the cumulative incidence in the state was 19.5% (15.9–23.7) and the cumulative undercount factor was 2.3 (1.9–2.9).

Ohio Governor Mike Dewine declared a state of emergency on March 9th 2020 and the state's first stay-at-home order took effect on March 23rd 2020. In mid-April, the Governor declared that businesses could begin to reopen on May 1st. Figure 2.3 shows that the first wave of infections, which picked up in March 2020 and likely peaked by late April 2020, did not die out but rather leveled out to a sustained spread through the summer of 2020. The posterior median of the reproductive number $r(t)$ in the state hovered around 1 from early April through mid-September and increased thereafter as the second wave of infections

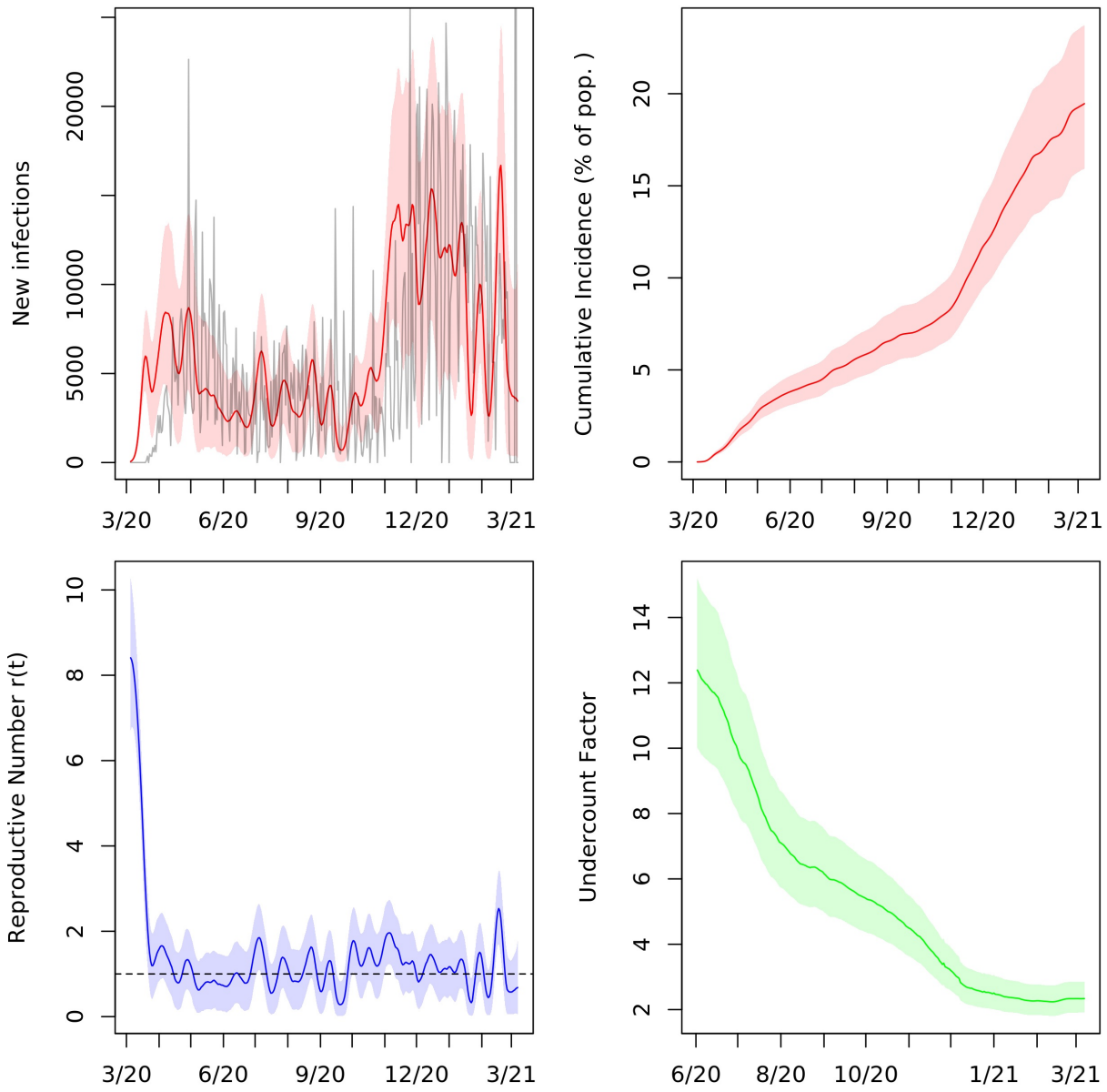


Figure 2.3: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, reproductive number $r(t)$, and cumulative undercount in Ohio from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

began in the fall.

2.3.3 Connecticut

We estimate an IFR of 1.37% (95% interval 1.10–1.70) in Connecticut. As of March 7, 2021, 15.9% (12.9–19.9) of the state's population has been infected, leading to an undercount factor of 2.0 (1.6–2.5).

According to our model as of April 26, 2020, 5.7% (4.6–7.1) of the state's population had recovered from COVID. In comparison, Havers et al. estimated a seroprevalence of 4.9% (95% interval 3.6–6.5) in the state in the period April 26–May 3 (Havers et al., 2020). Their study relied on a convenience sample of residual blood specimens collected for clinical purposes, and so it may have been affected by selection bias, as well as imperfect sensitivity and specificity of the antibody test used. Nevertheless, their estimate agrees well with the result from our model.

By July 5, 2020, our estimate of the recovered population increased to 8.9% (7.2–11.1). By comparison, in a random sample blood test survey, Mahajan et al. (2021) reported a seroprevalence of 4.0% (90% interval 2.0–6.0) for the period June 10–July 29, which is significantly lower. While our estimates disagree with those of Mahajan et al. (2021), we note that the survey response rate was low at 7.8%, raising the possibility of significant nonresponse bias. For this reason, we did not include the Connecticut survey as a source of data in our analysis.

2.3.4 New York

We estimate an IFR of 1.12% (95% interval 0.87–1.42) for New York state. As of March 7, 2021, 18.6% (14.7–23.9) of the state had been infected, yielding an undercount factor of 2.1 (1.7–2.8) through that date.

We know of no other estimates of the IFR in New York in the literature. However, Yang,

Kandula, et al. (2021) estimated an IFR of 1.39% (95% interval 1.04–1.77) for the first wave in New York City through June 6, 2020, based on available testing, mortality, and mobility data. According to New York City Health Department (2021) data, this period accounted for more than 85% of COVID deaths in the city and 57% of all confirmed COVID deaths (not including probable deaths) in the state through the first week of January 2021. As such, we expect the IFR for the state as a whole to have been similar to that of NYC during the spring of 2020, and our results are consistent with those of Yang, Kandula, et al. (2021).

We estimate that by June 6, 11.5% of the state's population (95% interval 9.0–14.7), or about 2.2 million people, had been infected with the novel coronavirus. Multiplying that number by the fraction of confirmed COVID deaths in the state occurring in NYC during that period yields 1.7 million infections, or 20% of the city's population. This number matches that of Stadlbauer et al. (2021), who measured 20% seroprevalence in NYC at that time based on randomly sampled residual plasma collected from patients at Mount Sinai Hospital scheduled for routine care visits unrelated to COVID-19.

2.3.5 United States

We summed posterior samples of the SIR trajectories from all the states to obtain estimates of viral prevalence in the United States on each day. The results are summarized in Figure 2.4. For each sampled trajectory of the infection curve, we calculated an effective contact parameter β_t for the entire country for each day from the SIR equations (2.1).

As of March 7, 2021, we estimate that 19.7% of the US population, or about 65 million people, had been infected with SARS-CoV-2. This suggests that the US was far from reaching herd immunity and that it was unlikely to do so from infections alone in the short term while state and local governments continue to implement lockdowns and other mitigations. Up to that date, we estimate that one out of every 2.3 infections in the US had been confirmed via testing. This implies that approximately 60% of all infections in the country had gone

unreported.

In the top left panel of Figure 2.4, which exhibits estimates of new infections on each day in the US, we plot reported COVID deaths per 1000 population shifted back 23 days (which is the mean of the time-to-death distribution τ). In the plot, we divide deaths per 1000 by 0.0068. This is the point estimate of IFR reported by Meyerowitz-Katz and Merone (2020) in their meta-analysis of 24 IFR estimates from a wide range of countries published between February and June 2020. The two curves have a substantial overlap, suggesting that the IFR implied by our estimates of true infections in the USA is consistent with their findings.

2.3.6 Implications for herd immunity

To illustrate the potential use of our method, we conducted a simulation to assess the implications of our results for herd immunity in the US. We project the SIR model for the US forward from January 6, 2021, and incorporate vaccine administration into the dynamics. We make the following strong assumptions:

1. Recovered individuals are immune to the virus, i.e., reinfection does not occur.

2. Immunity is conferred upon becoming fully vaccinated. Data tracking the number of people in the US fully vaccinated on each day are available from Our World in Data (Mathieu et al., 2020). Beyond April 16, 2021, we assume that the number fully vaccinated on each day follows the linear trend it had exhibited so far until reaching 2 million per day (see Figure 2.5(d)). After that, we assume that the number fully vaccinated per day remains at 2 million.

3. After January 6, the reproductive number $r(t)$ follows an AR(1) model with mean estimated from the sampled posterior trajectories of $r(t)$ through January 6.

The first point merits further discussion. Our projections that follow are particularly sensitive to this assumption. It may turn out that individuals who have been vaccinated

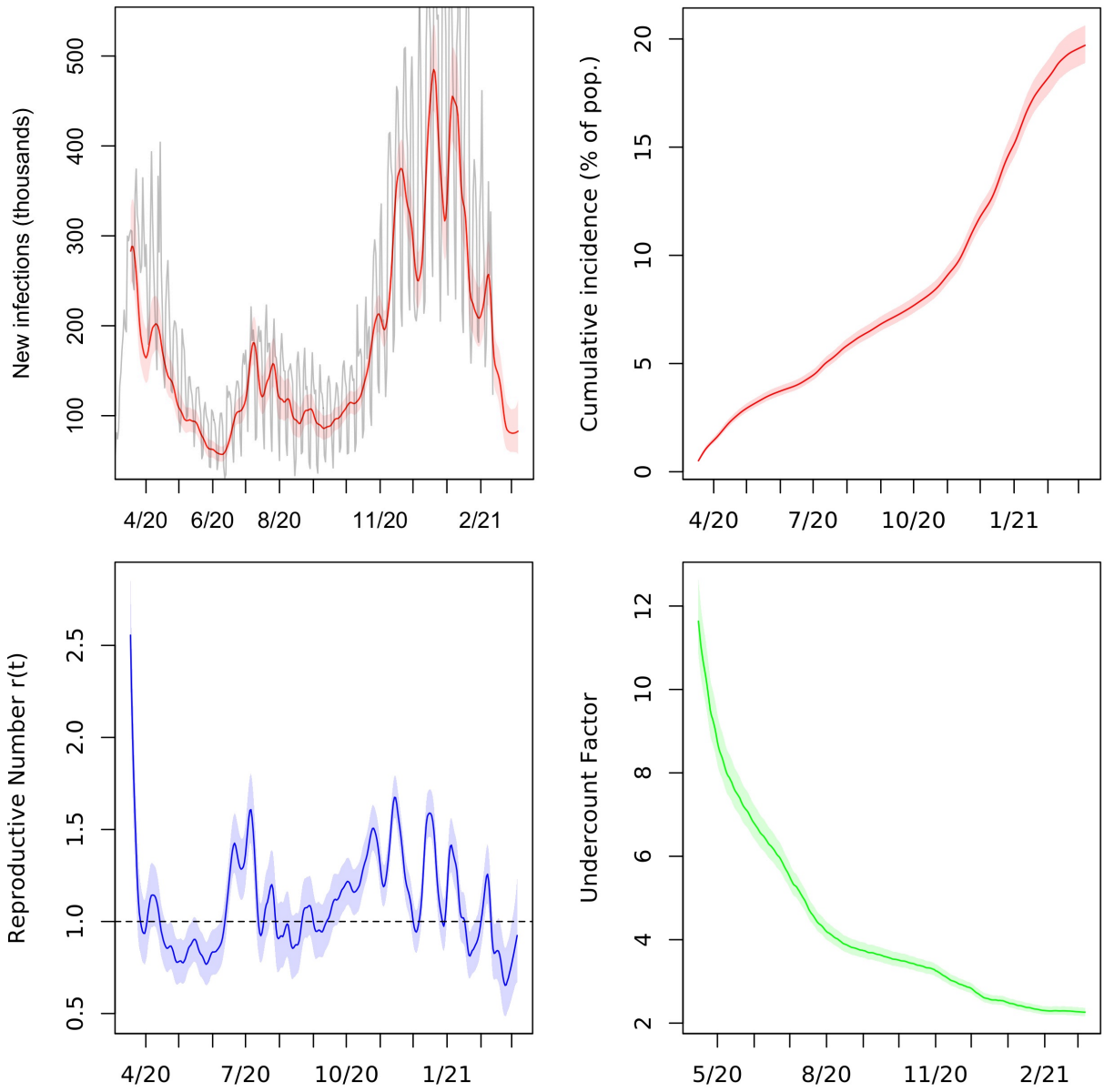


Figure 2.4: Aggregated estimates of new infections, cumulative incidence, reproductive number $r(t)$, and cumulative undercount for the United States from March 2020 to March 2021. In the top left panel, deaths (in thousands) divided by 0.0068 and shifted back 23 days are plotted in grey for comparison.

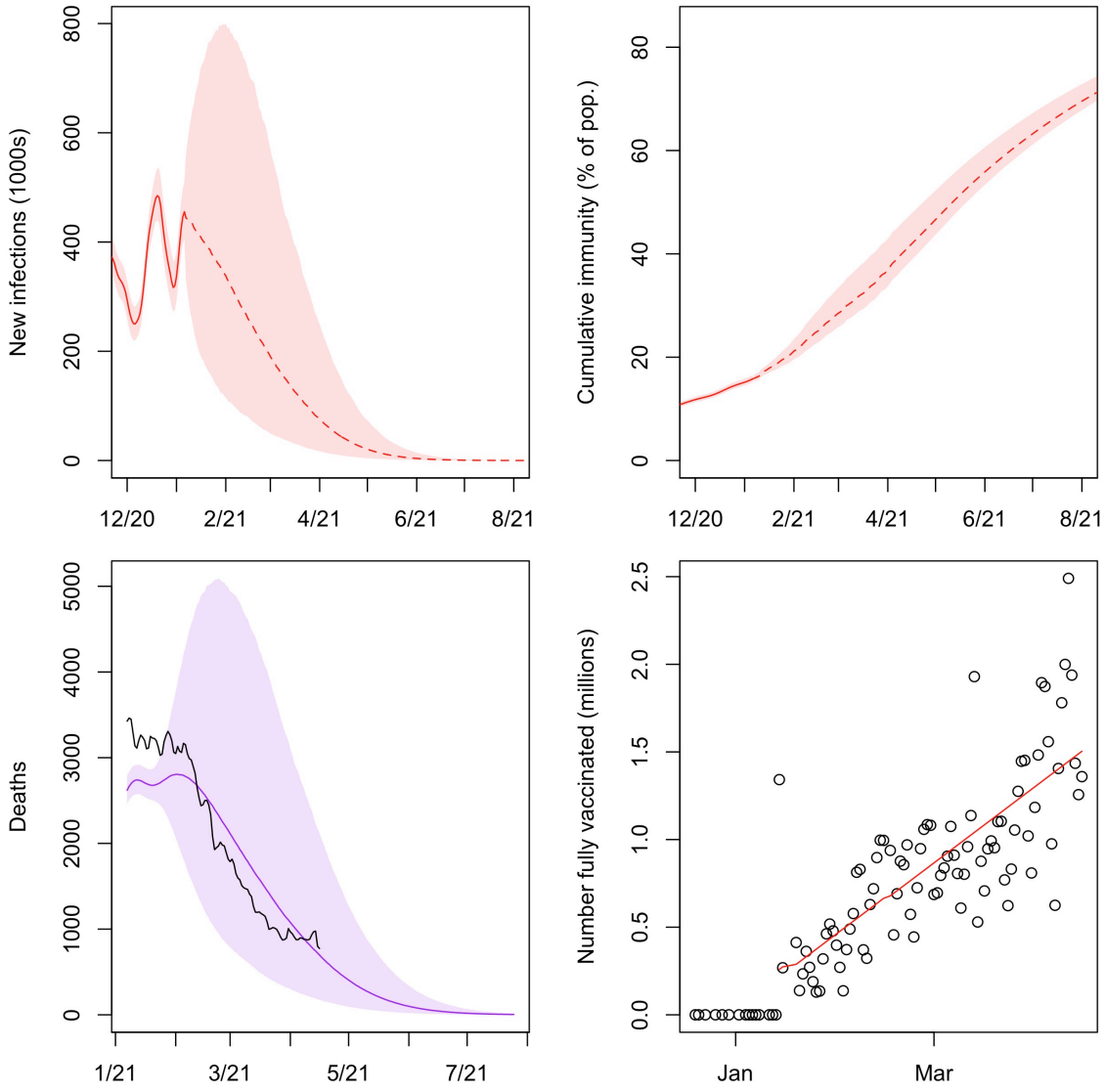


Figure 2.5: 95% credible intervals for the predictive distributions of new infections (a), cumulative immunity (viral incidence and full vaccinations) (b), and COVID deaths (c) in the US projected out from January 2021 through July 2021. In panel (c), the 7 day moving average of COVID deaths is plotted in black. (d) Scatterplot of the number of people newly fully vaccinated on each day in the United States as reported by Our World in Data (Mathieu et al., 2020). The line of best fit is plotted in red.

or previously infected are still susceptible to new variants of the virus that are cropping up and will continue to spread. It is also possible that the natural immunity conferred by asymptomatic and mild infections that elicited minimal immune response, which constitute a large portion of the total, will not last long enough to prevent widespread reinfection in the next few months. In either case, if Assumption 1 is violated then we may experience further waves of infection and delayed progress towards herd immunity.

Assumption 3 requires that the reproductive number oscillates around its mean, which is approximately 1.1 based on our estimates of $r(t)$ through January 6, 2021. This assumption is borne out by the plots of $r(t)$ in Figures 2.2, 2.3, 2.4, and for many other states (see plots in the Appendix). A possible explanation for this trend is the public and governmental response to deviations of $r(t)$ from 1. As $r(t)$ exceeds 1 and cases rise, lockdowns and other non-pharmaceutical interventions may be implemented to contain the virus; as $r(t)$ drops below 1 and cases dwindle, businesses, such as bars and restaurants, may be allowed to re-open, causing $r(t)$ to increase. When fitting the model in time periods for which we have data, we assumed instead that $r(t)$ follows a random walk, as described in the Materials and Methods section below. This process works well for estimation, but the variance can increase too much if used for projection beyond the short term. The stochastic autoregressive model captures future uncertainty in $r(t)$ more accurately. We also estimate the autocorrelation and variance parameters of the AR(1) model from the sampled trajectories of $r(t)$.

We project the 40,000 samples from the posterior distribution of the US infection trajectory forward under the modified SIR model described above. New infections and cumulative immunity (the percentage of the population previously infected or fully vaccinated) on each day are plotted in Figure 2.5. Based on our simulation, we find that the number of new infections per day in the country would likely fall below 5,000, about one hundredth of the winter peak, by June 2021, if our assumptions are valid. At this point, the virus' spread through the population will have been effectively suppressed. In getting there, we project

that we will incur another 18–31 million new infections, beginning from January 7. These numbers are obtained as the interquartile range of the projected cumulative incidence. Note that at that point, our model suggests that cumulative immunity will be 60% or less.

To put this in perspective, there were about 360,000 confirmed COVID deaths and 52 million infections (by our reckoning) as of January 6, 2021. Assuming an IFR of 0.68%, the additional 18–31 million new infections would lead to 186–270 thousand more COVID deaths, with 173–245 thousand occurring between January 7 and April 16, 2021. According to COVID data reported by The New York Times COVID-19 Data Team (2021), there were 204 thousand COVID deaths in the country between January 7 and April 16, which is consistent with our projections. Figure 2.5(c) also demonstrates that the predictive distribution of deaths from our projections matches up well with the data. We find that the projections given here are not very sensitive to plausible modifications of Assumptions 2 and 3.

2.4 Discussion

To craft and implement effective policy and mitigation strategies, policymakers need reliable assessments of the impact of previous non-pharmaceutical interventions on the transmission rate of the disease. We have developed a simple Bayesian model of the dynamics of SARS-CoV-2 transmission incorporating readily available time series data tracking the virus, as well as statewide representative point prevalence surveys conducted in Indiana and Ohio, which are the highest quality random testing surveys carried out to date. We present estimates of the infection fatality rate and the time-varying viral prevalence and reproductive number $r(t)$ in each US state on each day. Our results indicate that a large majority of COVID infections go unreported. Even so, we find that the US was still far from reaching herd immunity to the virus in early March 2021 from infections alone. This suggests that continued mitigation and an aggressive vaccination effort are necessary to surpass the herd immunity threshold without incurring many more deaths due to the disease. This work demonstrates the value

of random sample testing in response to this and future pandemics.

By incorporating testing and case data aggregated over any period of time, our additive model for positive tests in equation (2.2) allows us to avoid using data at the daily level, which can be very unreliable. For example, the reported cumulative number of tests administered in a state may not be updated for up to two weeks at a time, or it may decrease from one day to the next as data are deduplicated upon further review. The latter scenario frequently occurs with reported cases as well. Working with data at the daily level generally requires using some kind of moving average, which washes out stochasticity in the data and leads to oversmoothing inconsistent with the high overdispersion of SARS-CoV-2 transmission (Endo et al., 2020).

Our inference relies on daily reported deaths due to COVID in each state as opposed to excess deaths. Because of the possibility of death misclassification, excess death data represent a mix of confirmed COVID deaths and deaths from other causes. Nevertheless, relying on reported deaths is a potential source of bias, as they are affected by the accuracy of cause-of-death determinations. Their numbers can fall significantly below excess death counts and may undershoot the true number of deaths due to the disease (National Academies of Sciences, Engineering, and Medicine, 2020). Ascertainment of COVID deaths may vary between states, with the cumulative excess death count since the start of the pandemic exceeding reported COVID deaths by upwards of 50% in some states, according to a New York Times analysis of CDC mortality data (Katz et al., 2021). Consequently, our results may underestimate viral incidence in those states.

CDC (2021) estimated a total of 83 million infections in the US through December 2020, which is substantially larger than our estimate of 50 million infections in that period. Their numbers are based on the work of Reese et al. (2020), who infer COVID incidence in the US using a multiplier model to account for under-detection in the number of confirmed cases. Beyond the limitations of our study discussed above, there are a few possible explanations

for the difference in our estimates. Reese et al. (2020) base their estimates on nationally reported laboratory-confirmed cases, which do not constitute a probabilistic sample of the population. To this point, the authors remark that "...some infections, such as those among healthcare workers or from outbreaks in congregate residential settings, may be more likely to be tested and nationally reported compared with the general population, and could overestimate nonhospitalized cases and infections". Furthermore, the multiplier in their model relies on documented rates of test administration and care-seeking among symptomatic COVID patients. Reese et al. (2020) note that data on rates of test administration in this group are limited, especially at the local level. As such, Reese et al. (2020) do not account for geographic variation in testing, which is a potential source of bias.

Chapter 3

EVALUATING AND OPTIMIZING NON-PHARMACEUTICAL INTERVENTIONS TO COMBAT INFECTIOUS DISEASE

Non-pharmaceutical interventions (NPIs), which were governments' primary tools to mitigate SARS-CoV-2 transmission prior to the arrival of COVID-19 vaccines and antiviral treatments, necessitated a trade-off between the health impacts of viral spread and the social and economic costs of restrictions. To address this trade-off in a principled way, we develop a statistical decision framework and conduct a cost-effectiveness analysis of NPI policies enacted at the state level in the United States in 2020. We quantitatively evaluate the efficacy and gross impacts of the policy schedules implemented during the pandemic and derive optimal cost-effective NPI strategies. Although school closures significantly reduced viral transmission, their social impact in terms of student learning loss was far too costly. Conditional on the other policies enacted, extended school closures imposed a cost to the nation's youth in service of its older generations, reducing the latter's risk of death at the expense of $2 trillion (USD2020) in future GDP. Moreover, we find that this marginal trade-off between school closure and COVID deaths was not inescapable: more timely, stringent, and enduring use of other measures in combination would have sufficed to maintain similar or lower mortality rates without incurring profound learning loss. Optimal policies involve consistent implementation of mask mandates, public test availability, contact tracing, social distancing orders, and reactive workplace closures, with no closure of schools beyond the usual 16 weeks of break per year. Their use would have reduced the gross impact of the pandemic in the U.S. in 2020 from $5.1 trillion to $2.4 trillion and, with high probability, saved lives. In sensitivity analysis, we assess the robustness of our results to plausible parametric

variation. We discuss the implications of our findings and methodology for the implementation of NPIs in other contexts and the use of incremental cost-effectiveness ratios (ICERs) in infectious disease research.

3.1 Introduction

In the year prior to the arrival of COVID vaccines and other pharmaceutical interventions, non-pharmaceutical interventions (NPIs)—including school and workplace closures, social distancing, masking, testing, and contact tracing—were the primary tools for mitigating the spread of SARS-CoV-2. The use of NPIs posed significant challenges to decision-makers at every level of government, who were forced to make difficult and consequential real-time decisions with limited data and amidst contentious political debate (Adolph et al., 2021). While they substantially reduced viral transmission, extended national and sub-national lockdowns had severe deleterious social and economic consequences globally—including disrupted economic output, job loss, and student learning loss (Chetty et al., 2024; UNICEF et al., 2021)—on top of the already staggering health impacts of the pandemic. To address this trade-off in a principled way, we develop a statistical decision framework and conduct a cost-effectiveness analysis of non-pharmaceutical intervention policies enacted at the state level in the United States in 2020.

Our analysis is composed of three steps. We first build a Bayesian mechanistic epidemiological model estimating SARS-CoV-2 prevalence and transmission rates in each state over time based on prior work leveraging random sample testing surveys to debias clinical COVID data (Irons & Raftery, 2021). We next estimate the effects of NPIs on viral transmission in all states jointly using a Bayesian hierarchical regression model controlling for temporal autocorrelation and endogenous behavioral responses linked to fear of infection. Finally, we couple these estimates with monetary costs associated to the social, economic, and health consequences of infection and NPIs drawn from the COVID literature in order to quantita-

tively evaluate the efficacy and gross impacts of the policy schedules implemented during the pandemic and to derive strategies that optimally navigate the trade-off between restrictions and viral spread.

Although school closures significantly reduced viral transmission, their social impact in terms of student learning loss was far too costly. Conditional on the other policies enacted, extended school closures imposed a cost to the nation's youth in service of its older generations, reducing the latter's risk of death at the expense of $2 trillion (USD2020) in future GDP.[1] Moreover, we find that this marginal trade-off between school closure and COVID deaths was not inescapable: more timely, stringent, and enduring use of other measures in combination would have sufficed to maintain similar or lower mortality rates without incurring profound learning loss.

Optimal NPI strategies involve consistent implementation of mask mandates, public test availability, contact tracing, social distancing orders, and reactive workplace closures, with no closure of schools beyond the usual 16 weeks of break per year. Their use would have reduced the gross impact of the pandemic in the U.S. in 2020 from $5.1 trillion to $2.4 trillion and, with high probability, saved lives. In any case, these impacts—health, economic, and social—were felt disproportionately by marginalized populations (Betthäuser et al., 2023; Chetty et al., 2024; Tai et al., 2021). In sensitivity analysis, our conclusions are robust to plausible parametric variation, except that: the optimal duration of workplace closures decreases as their effect on unemployment rises; and the optimal policy involves some amount of school closure when we assume both that the average COVID infection is extremely costly—based on a value per statistical COVID death (VSCD) equal to the value of a statistical life (VSL) (conventionally about $11 million in USD 2020), which does not adjust for the age profile of COVID mortality—and that school closures are relatively cheap—based on a cost of student

[1]This number and those that follow are based on the cost of learning loss reported by Psacharopoulos et al. (2021), which is a conservative estimate. The cost of learning loss estimated by Hanushek and Woessmann (2020) is much higher. We use the results of Psacharopoulos et al. (2021) as a baseline.

learning loss that assumes distance learning is 90% as effective as in-person schooling.

While our study focuses on COVID-19 in the U.S. prior to the arrival of vaccines, our qualitative findings shed light on NPI implementation in other settings. Masking, testing, and tracing are relatively cheap and likely to remain cost-effective universally: for severe and relatively mild pandemics; in lower resource settings; and after effective pharmaceutical interventions become available. After the arrival of vaccines and antiviral treatments, workplace closures and social distancing measures should be enacted more sparingly. Although school closures were not cost-effective, evidence suggests that distance learning helped to mitigate learning loss. Consequently, extended school closures are likely to be relatively more costly in low- and middle-income countries with younger populations and less capacity to provide effective education remotely (UNICEF et al., 2021). Likewise, with fewer opportunities for remote work and less online economic activity, workplace closures, stay-at-home orders, and other social distancing measures may be more costly in these countries (Barnett-Howell et al., 2021; Bloom et al., 2022; Decerf et al., 2021). For less virulent diseases with a similar age pattern of death, extended school closures should never be implemented and extended workplace closures and social distancing measures should be mandated with care. If possible, more targeted interventions should be utilized (Acemoglu et al., 2021).

Related literature. The literature studying non-pharmaceutical interventions in response to COVID-19 and past pandemics is vast. Bloom et al. (2022) and Brodeur et al. (2021) provide reviews of the relevant economics literature. Our work is most closely related to: studies estimating associations and inferring causal effects of NPIs on viral transmission (Banholzer et al., 2021; Bo et al., 2021; Brauner et al., 2021; Cauchemez et al., 2008; Chernozhukov et al., 2021; Flaxman et al., 2020; Haug et al., 2020; Hsiang et al., 2020; Karaivanov et al., 2021; Li et al., 2021; Liu et al., 2021; Sharma et al., 2021); studies quantifying the gross health and economic impacts of pandemics and the associated policy response (Cutler & Summers, 2020; Ferguson et al., 2020; Greenstone & Nigam, 2020; Hall et al., 2020; Kaplan

et al., 2022; Prager et al., 2017; Thunström et al., 2020); and modeling papers studying the (optimal) control of epidemics and the cost-effectiveness of non-pharmaceutical interventions appearing in the economics (Acemoglu et al., 2021; Adda, 2016; Alvarez et al., 2021; Barrot et al., 2024; Eichenbaum et al., 2021; Farboodi et al., 2021; Gollier, 2020; Jones et al., 2021; Krueger et al., 2022; Maharaj & Kleczkowski, 2012; Reluga, 2010) and public health literature (Brown et al., 2011; Dauelsberg et al., 2024; Ferguson et al., 2006; Halder et al., 2011; Halloran et al., 2008; Kelso et al., 2013; Keogh-Brown et al., 2020; Keogh-Brown et al., 2010; Milne et al., 2013; Perlroth et al., 2010; Sander et al., 2009; Smith et al., 2009; Xue et al., 2012). Throughout the text, we highlight how the results of our models compare to these studies.

Our methodology builds upon and extends the literature studying NPI mitigation of infectious disease in order to address a number of gaps that severely limit its value in informing and evaluating policies. Firstly, to the best of our knowledge, the optimal control of a pandemic has never been studied using statistical decision theory. We take a data-driven approach to the NPI decision process by estimating and accounting for uncertainty in key parameters, including viral prevalence, reproduction numbers, and the effects of NPIs and other endogenous and (unpredictable) exogenous factors on transmission rates. In particular, we produce probabilistic estimates of SARS-CoV-2 prevalence over time, which are necessary to properly account for the magnitude and uncertainty of costs associated to infections. As another consequence of our data-driven approach, our model is able to capture the complex and stochastic temporal trends of SARS-CoV-2 transmission (e.g., multiple waves, superspreader events, the introduction of new infections via travel, and random fluctuations) that are missed by studies relying on stylized deterministic epidemiological simulation models in which long-run future dynamics can be predicted with perfect accuracy. This allows us to define and evaluate realistic counterfactual scenarios under different NPI policies conditional on what was actually observed during the pandemic. Given the largely unpredictable nature

of SARS-CoV-2 transmission, we find that the structure of the optimal NPI strategy is remarkably simple and consistent across time and space, with the planner required to respond to COVID dynamics in real time to a minimal degree.

Secondly, we model the costs and effects on viral transmission of multiple specific NPIs calibrated to U.S. economic and state-level COVID data. Modeling papers studying the cost-effectiveness of NPIs and their optimal control of pandemics often consider a limited toolkit. They tend to focus on a minimal collection of interventions, often a single NPI (Barrot et al., 2024; Brown et al., 2011; Dauelsberg et al., 2024; Xue et al., 2012), which, in the economics literature, usually represents a catch-all "social distancing", "containment", or "lock-down" policy (Acemoglu et al., 2021; Alvarez et al., 2021; Eichenbaum et al., 2021; Farboodi et al., 2021; Gollier, 2020; Jones et al., 2021; Krueger et al., 2022; Maharaj & Kleczkowski, 2012; Reluga, 2010). This is impractical, as we generally have a range of tools at our disposal, and NPIs are known to be more effective—and, therefore, more cost-effective—in combination (Ferguson et al., 2006; Halder et al., 2011; Juneau et al., 2022; Kelso et al., 2013; Milne et al., 2013; Perlroth et al., 2010). Furthermore, if we consider only a single instrument, we may erroneously conclude that its implementation is cost-effective because we implicitly assume that other policies either cannot be enacted or cannot be modified from a set schedule. As we discuss in greater detail in Section 3.3.3, the cost-effectiveness of any single intervention is highly context-specific and, in particular, depends strongly on the other policy options at hand. We study a suite of 11 non-pharmaceutical interventions. As such, our findings are more precise and informative than the broad qualitative guidance drawn from prior studies in the context of COVID-19—e.g., that "lock-down" is cost-effective and optimal when implemented early and stringently (Alvarez et al., 2021). By evaluating a range of NPIs, we can disaggregate policies to conclude that testing, tracing, masking, reactive workplace closure, and social distancing measures (not including extended school closure) combine to form an optimal cost-effective strategy.

Thirdly, we depart from the large majority of studies assessing the economic impact and cost-effectiveness of school closure during pandemics by factoring in costs associated to student learning loss (Brown et al., 2011; Dauelsberg et al., 2024; Deb et al., 2022; Halder et al., 2011; Kelso et al., 2013; Keogh-Brown et al., 2020; Keogh-Brown et al., 2010; Lempel et al., 2009; Milne et al., 2013; Perlroth et al., 2010; Sadique et al., 2008; Sander et al., 2009; Smith et al., 2009; Verschuur et al., 2021; Viner et al., 2020; Walmsley et al., 2023). Most studies in this body of literature quantify the total cost of school closure as a sum of direct costs arising from: lost productivity of teachers and school staff; and workplace absenteeism of parents or childcare costs resulting from students staying home. However, the indirect costs of school closure are substantial. Students suffering acute learning loss go on to become less skilled and less productive members of the workforce, which in turn leads to future losses in personal income and national GDP (Hanushek & Woessmann, 2020). We account for the net present value of these amortized future losses to society, which can be staggering in magnitude, based on estimates of the cost of learning loss from the education economics literature (Azevedo et al., 2021; Hanushek & Woessmann, 2020; Psacharopoulos et al., 2021) and recent estimates of the amount of learning loss accrued during COVID-19 school closures (Betthäuser et al., 2023). Considering other indirect costs, we note that: school disruptions and decreases in educational attainment may be associated with various negative health outcomes among students, including depression, anxiety, and decreased life expectancy (Christakis et al., 2020; Viner et al., 2022);[2] and school closures can cause significant healthcare worker absenteeism, potentially negating some or all of the mortality benefits from school-closure-related reductions in SARS-CoV-2 transmission (Bayham & Fenichel, 2020; Lempel et al., 2009). Regarding the former, we do not account for potential

[2]Similarly, there is some evidence that COVID-19 restrictions are associated with increases in drug overdose fatalities (Wolf et al., 2024). Nevertheless, as with school closures (and for the same reasons), we do not take into account potential physical and mental health costs related to other NPI policies. We discuss this further in Section 3.4.

downstream physical and mental health costs of school closures as comprehensive causal links and quantitative estimates have not been established. Regarding the latter, we do not account for health impacts related to healthcare personnel absenteeism as quantifying their cost is challenging. For these and other reasons, which we discuss in more detail in Section 3.2.4, we believe that our accounting of the costs of school closure is conservative.

In our review of the literature, we found only two cost-benefit analyses of school closure that account for learning loss (Adda, 2016; Xue et al., 2012). Studying pandemic flu, Xue et al. (2012) find that school closures are not cost-effective for mild strains (such as the 2009 H1N1 virus), but they generate net benefits in the context of more severe pandemics, such as the 1918 Spanish flu. Similarly, studying historical outbreaks of influenza, gastroenteritis, and chickenpox in France, Adda (2016) determines that school closures were not cost-effective, but that they would become beneficial for slightly more lethal epidemics. Notably, Xue et al. (2012) model school closure in isolation, i.e., they do not consider the availability of other interventions, and Adda (2016) considers only school and public transport closure. Their results concur with a number of other studies (not accounting for learning loss) finding that extended school closures are cost-effective for severe pandemics (Dauelsberg et al., 2024; Deb et al., 2022; Kelso et al., 2013; Milne et al., 2013; Perlroth et al., 2010). To the contrary, we demonstrate that COVID-19 school closures were not cost-effective.[3]

Additional methodological contributions of our approach include a novel zero-inflated negative binomial model that flexibly captures well-known reporting idiosyncrasies and overdispersion in clinical COVID data. As a result, our method eliminates the need for *ad hoc* data cleaning and smoothing procedures that can complicate the analysis pipeline, yield poorly calibrated prediction intervals, and potentially bias transmission rate estimates based

[3]While we find that extended school closure is not cost-effective, a relevant (and potentially cost-effective) counterfactual would have been a reactive school closure of limited duration at the beginning of the pandemic (i.e., spring of 2020) compensated by an extended school year stretching into the summer. Such a strategy would not have incurred student learning loss; it would have merely shifted the summer break toward spring to allow for an urgent response to the initial outbreak.

on over-smoothed data. Furthermore, we implement a two-stage modeling procedure that first estimates the time-varying effective reproduction number in each US state individually, followed by a joint hierarchical model across states that estimates pooled effects of NPIs on transmission dynamics. This approach allows for efficient Bayesian computation by parallelizing model fits across states. While our results are specific to the COVID-19 pandemic, our methods can be used more widely to evaluate public health interventions against infectious disease.

Outline of the text. Section 3.2 details our methodology, including specifics of the data and implementation, the construction of our models, and the elicitation of costs associated to infections and NPIs. Section 3.3 reports the baseline results of our models and sensitivity analysis, compares our results to others in the literature, and discusses implications of our methodology and results for the use of incremental cost-effectiveness ratios (ICERs) in infectious disease. Section 3.4 provides concluding remarks and discusses qualifications and limitations of our methodology.

3.2 Methods

3.2.1 Data and implementation

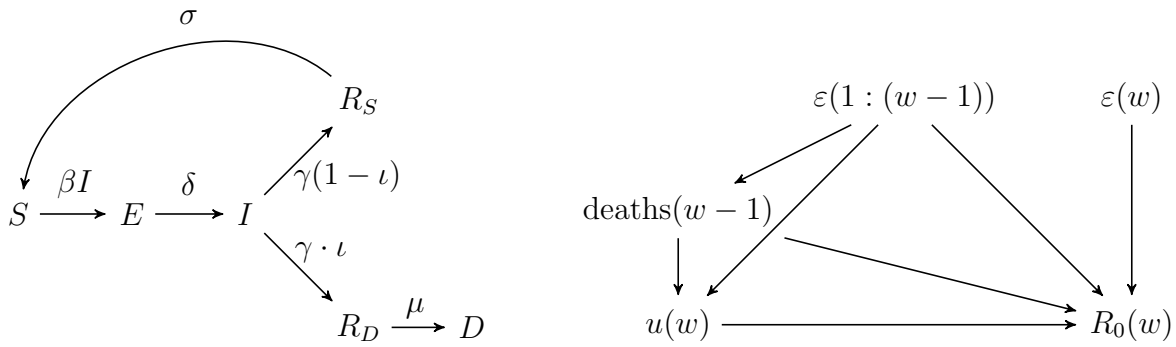
We obtained U.S. state-level daily counts of confirmed COVID cases and deaths in 2020 from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) (Dong et al., 2020). If a negative number of deaths or cases were reported on a given day—often due to retroactive changes in the reported cumulative death or case count for record deduplication or changes in data reporting by the state government—we assume that the cumulative death (case) count on that day was the correct one and set the number of deaths (cases) incident on prior days to zero until the overall cumulative count is non-decreasing. We begin modeling viral transmission in each

state on the first day on which more than one death or case is reported.

We obtained state-level government NPI policies reported daily from the Oxford COVID-19 Government Response Tracker (OxCGRT) (Hale et al., 2021). In converting the ordinal policy levels to numerical values, we followed OxCGRT's methodology for calculating indices, in which ordinal levels are equally spaced numerically and a targeted (as opposed to general) intervention is treated as a half-step between ordinal levels. We rescale each policy value to lie between 0 and 1, with 1 denoting the most stringent policy. If a policy is not recorded on a given day, we set its value to that on the previous day on which the policy was recorded, or we set it to zero if at the beginning of the study period. We average daily policy values at the weekly level for our NPI regression model.

We obtained state-level counts of SARS-CoV-2 PCR tests administered on each day from the COVID Tracking Project (The Atlantic, 2021). We obtained daily average surface temperature data for the largest city in each state using the Meteostat Python package (Meteostat, 2022).

Data cleaning was conducted in R and Python. All models were fit in R using the CmdStanR package, with MCMC convergence assessed using the diagnostics provided therein (Gabry et al., 2024; R Core Team, 2020). We used the optimParallel R package for NPI policy optimization (Gerber & Furrer, 2019). To determine the optimal NPI strategy in each state based on the cost function outlined in Section 3.2.4, we used a combination of 8 random and hand-specified parameter initializations and kept the policy yielding the smallest value of the cost function. In practice, we find that the results of the optimization are robust to the initial parameters, which is reflected in our results. Code to reproduce our analysis is available at the GitHub repository (https://github.com/njirons/covidOC).



(a) SEIRD model.

(b) NPI regression model DAG.

Figure 3.1: Graphical models.

3.2.2 Bayesian epidemiological model

We begin by describing the Bayesian epidemiological model used to estimate SARS-CoV-2 prevalence and transmission rates in each US state in 2020. In our two-stage estimation procedure, we first fit this epidemiological model to each state separately. Next, the time-varying basic reproduction numbers $R_0(t)$ output by the model in each state are fed into the Bayesian hierarchical regression model described below in Section 3.2.3 in which we jointly model transmission rates in all states as a function of NPIs.

SEIRD model

Our discrete-time Bayesian Susceptible-Exposed-Infectious-Removed-Deceased (SEIRD) model builds on the model of Irons and Raftery (2021), which was used to estimate state-level SARS-CoV-2 prevalence in the first year of the pandemic based on reported cases, deaths, tests, and random testing surveys.

For a given US state, let $S(t)$ denote the proportion of susceptible people in the state on day t, $E(t)$ the proportion exposed but not yet infectious, $I(t)$ the proportion infectious, $R_S(t)$

the proportion recovered (survivors no longer infectious), $R_D(t)$ the proportion no longer infectious who will eventually succumb to the disease, and $D(t)$ the proportion decedent. These quantities evolve in time according to the equations

$$\begin{cases} S(t+1) - S(t) & = \sigma R_S(t) - \beta(t) S(t) I(t) \\ E(t+1) - E(t) & = \beta(t) S(t) I(t) - \delta E(t) \\ I(t+1) - I(t) & = \delta E(t) - \gamma I(t) \\ R_S(t+1) - R_S(t) & = \gamma(1-\iota) I(t) - \sigma R_S(t) \\ R_D(t+1) - R_D(t) & = \gamma \cdot \iota I(t) - \mu R_D(t) \\ D(t+1) - D(t) & = \mu R_D(t). \end{cases} \tag{3.1}$$

A graphical model of this process is depicted in Figure 3.1a.

Members of the population move from susceptible to exposed after contact with an infectious person with rate $\beta(t)$, which is allowed to vary in time to account for variation in exposure due to social distancing and other factors. Following the incubation period (with duration δ^{-1}), exposed people become infectious and are subsequently removed at rate γ, at which point they no longer infect others. A proportion ι (the infection fatality rate, or IFR) of removed individuals die from COVID at temporal rate μ or remain alive and eventually lose acquired immunity, thereby becoming susceptible again at rate σ. As a simplifying approximation, our model assumes a conserved population, i.e., there are no births and no deaths due to competing risks:

$$S(t) + E(t) + I(t) + R_S(t) + R_D(t) + D(t) = N$$

for all times t, where N is the state's total population. Note that the time-varying basic reproduction number $R_0(t)$ and effective reproduction number $R_e(t)$, which describe rates of transmission in the initial and current population, respectively, are given by $R_0(t) = \beta(t)/\gamma$ and $R_e(t) = S(t)R_0(t)$.

We assume that γ^{-1}, the average length in days of the infectious period, is determined by the disease and constant over time. We make the same assumption for the other biological parameters introduced above. In particular, while the IFR ι can realistically change over time, e.g., due to vaccination, the time period of our study focuses on viral transmission prior to widespread vaccine administration and circulation of novel SARS-CoV-2 strains with differential virulence. Estimates of the IFR over time in England based on regular random testing of the population found that, while the IFR did fluctuate in 2020, it hovered around 0.67% (Eales et al., 2023). This is consistent with the IFR estimated in a systematic meta-analysis in 2020 (Meyerowitz-Katz & Merone, 2020) and with the results of Irons and Raftery (2021).

Regarding prior specification, $R_0(t)$ is given a random walk structure on the log scale using the truncated log-normal distribution. We assume that $R_0(t)$ is constant during each week and, in an abuse of notation, write $R_0(w(t))$ to mean the value of R_0 in week $w(t)$ to which day t belongs. We have

$$\begin{aligned} R_0(w+1) &= \text{LogNormal}_{[0,R_0^{\max}]}(R_0(w), \sigma_R^2), \\ R_0(0) &\sim \text{Uniform}(0, R_0^{\max}), \\ \pi(\log \sigma_R^2) &\propto 1. \end{aligned}$$

We place a flat improper prior on the log-transformed scale parameter $\log \sigma_R^2$. We take $R_0^{\max} = 6.5$ to be the upper bound for the transmission rate based on (Liu et al., 2020). We

place a flat Dirichlet prior on the initial SEIRD components:

$$\begin{aligned}
S(0) &= (1-p) + p \cdot x_0(S), \\
E(0) &= p \cdot x_0(E), \\
I(0) &= p \cdot x_0(I), \\
R_S(0) &= p \cdot (x_0(R) + x_0(D))(1-\iota), \\
R_D(0) &= p \cdot x_0(R) \cdot \iota, \\
D(0) &= p \cdot x_0(D) \cdot \iota, \\
(x_0(S), x_0(E), x_0(I), x_0(R), x_0(D)) &\sim \text{Dirichlet}(1,1,1,1,1),
\end{aligned}$$

Here $p = 0.03$ is the upper bound on the proportion of the population potentially infected at or before time 0 (the first day of the study period). The remaining parameters are detailed in Table 3.1. We use state-specific IFR estimates using the posterior median reported in Irons and Raftery (2021).

Likelihood on deaths

In a given U.S. state, let $d(t)$ and $c(t)$ denote the number of COVID deaths and cases recorded in the state on day t, as recorded the JHU CSSE (Dong et al., 2020). To account for measurement error, idiosyncratic reporting, and overdispersion in viral transmission (Endo et al., 2020; Hasan et al., 2020; Kremer et al., 2021; Lau et al., 2020; Sneppen et al., 2021; Sun et al., 2021), we use a zero-inflated negative binomial model on $d(t)$ and $c(t)$. Many states inconsistently reported cases and deaths, often taking breaks over weekends and holidays, resulting in numerous spurious zeros in the data. We address this by assuming that any deaths or cases occurring on such a day are reported on the first subsequent day of accurate reporting.

Specifically, let Z_t indicate the event that the number of deaths occurring on day t is incorrectly reported as 0. We assume that the Z_t are independent and identically distributed

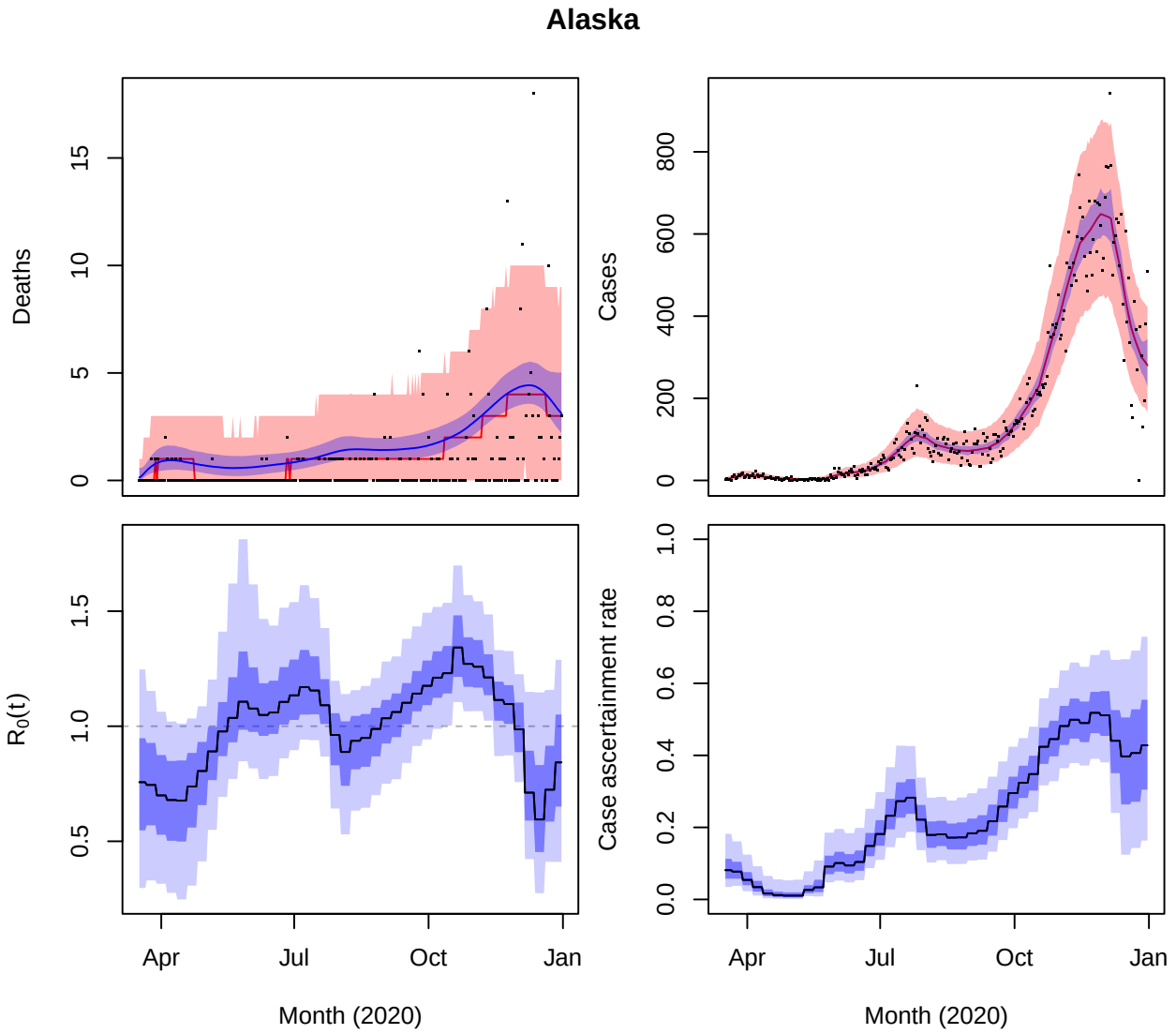


Figure 3.2: SEIRD model fit to COVID data in Alaska. **Top panels:** observed deaths $d(t)$ and cases $c(t)$ are plotted in black. Median and 90% credible intervals of the posterior predictive distributions of $d(t)$ and $c(t)$ are in red. Posterior median and 90% credible intervals of the underlying mean parameters $m_D(t)$ and $m_C(t)$ are in blue. **Bottom panels:** Posterior median, 50%, and 90% credible intervals for the basic reproduction number $R_0(t)$ and the case ascertainment rate CAR(t).

Table 3.1: Epidemiological parameters. All times are in days.

Parameter	**Value**	**Reference**
σ^{-1}: Mean duration of acquired immunity	270	Hall et al. (2022) Helfand et al. (2022) Pooley et al. (2023)
δ^{-1}: Mean duration of incubation period	5.1	Linton et al. (2020) Lauer et al. (2020b) Gallo et al. (2020) Wu et al. (2022)
γ^{-1}: Mean duration of infectious period	8.5	Gallo et al. (2020) Byrne et al. (2020)
μ^{-1}: Mean time to death after removal	10	Linton et al. (2020) Byrne et al. (2020)
τ_D: Mean time from exposure to death	23.6	Ward and Johnsen (2021)
ι_s: State-specific IFR	Varying (0.2–1.7%)	Irons and Raftery (2021)
$R_0^{\max}$: Upper bound on $R_0(t)$	6.5	Liu et al. (2020)
Mean time from case reporting to death	8.053	Jin (2021)
Standard deviation in time from case reporting to death	4.116	Jin (2021)

with $P(Z_t = 1) = \theta_D$. We know that $Z_t = 0$ on days with reported deaths ($d(t) > 0$) and our model conditions on this knowledge. Assume $t_0 < t_0 + k$ are days with $d(t_0) > 0$ and $d(t) = 0$ for all $t \in (t_0, t_0 + k]$. Note that some of the zeros on days $t \in (t_0, t_0 + k]$ could be due to misreporting, whereas others could be accurate reporting days on which zero deaths actually occurred. We marginalize over the unknown random variables Z_t for $t \in (t_0, t_0 + k]$ conditional on the assumptions that: the reported deaths $d(t_0)$ are centered at $m_D(t) = N\mu R_D(t_0)$, the true number of deaths on day t_0; in expectation, any deaths occurring on misreporting days are reported on the next day of accurate reporting. Under these assumptions, the underlying mean of observed cases $d(t_0 + k)$ conditional on $Z_{t_0+k} = 0$ is, after marginalizing over $Z_t, t \in (t_0, t_0 + k)$,

$$\begin{aligned} m_D(t_0 + k) &= \mathbb{E}[d(t_0 + k) | Z_{t_0+k} = 0] \\ &= \mathbb{E}\left[m_D(t_0 + k) + \sum_{t \in (t_0, t_0+k)} Z_t m_D(t) \right] \\ &= \sum_{t=0}^{k-1} m_D(t_0 + k - t)\theta_D^t. \end{aligned}$$

The likelihood on observed deaths is then given by the following zero-inflated negative binomial:

$$P\left(d(t) = d \mid m_D(t), \kappa_D(t), \theta_D\right) = \begin{cases} \theta_D + (1 - \theta_D) \cdot \text{NegBin2}(0, \kappa_D(t)^{-1}), & d = 0, \\ (1 - \theta_D) \cdot \text{NegBin2}(m_D(t), \kappa_D(t)^{-1}), & d > 0, \end{cases} \tag{3.2}$$

where $\text{NegBin2}(\mu, \tau)$ is parametrized to have mean μ and variance $\mu + \mu^2/\tau$. We allow the overdispersion parameter $\kappa_D(t)$ to depend on the mean as follows:

$$\kappa_D(t)^{-1} = \kappa_D^{-1} \left(\zeta_D m_D(t) + (1 - \zeta_D) \right),$$

where $\zeta_D \in [0, 1]$ is a proportion parameter and $\kappa_D \in (0, \infty)$. We use $\text{Uniform}(0, 1)$ priors on θ_D and ζ_D and a flat improper prior $\pi(\log \kappa_D) \propto 1$. Finally, with $\tilde{d}(0)$ representing the

cumulative deaths reported prior to the start of the modeling window, we use the likelihood

$$\tilde{d}(0) \sim \text{Poisson}(N \cdot D(0)).$$

This model flexibly interpolates between a count distribution with a linear mean-variance relationship (as with the overdispersed Poisson) when $\zeta_D = 0$ and a quadratic mean-variance relationship (as with the usual negative binomial) when $\zeta_D = 1$. We found that this modification was necessary to accurately capture dispersion in clinical data across a range of states. In some states, such as Ohio and Indiana, a standard Poisson was sufficient to produce well-calibrated posterior predictive distributions. In most other states, such as Texas and Florida, a negative binomial was required. Finally, there were some states, such as New York, in which Poisson predictive intervals were too narrow and negative binomial intervals were too wide, while predictive intervals derived from the model (3.2) were much better calibrated. Our model (3.2) can handle all of these cases. Figure 3.2 demonstrates the model's fit to COVID data in Alaska, which exhibit large and time-varying patterns of overdispersion and zero-inflation. The mean $m_D(t)$ of the likelihood on deaths provides a much smoother representation of the data and depicts more consistent trends in transmission reflected in the case data.

Likelihood on cases

Let $\nu(t) = N\beta(t)S(t)I(t)$ denote the number of new infections in the state on day t. We relate the true prevalence $\nu(t)$ to the number of cases $c(t)$ reported on each day using a compartmental model that accounts for time-varying imperfect case ascertainment and delays between exposure and case confirmation via testing. We define a "number of infections waiting to be confirmed" compartment $I_C(t)$ satisfying

$$I_C(t+1) = I_C(t)(1-\tau) + \text{CAR}(t+1)\nu(t+1),$$

where τ^{-1} is the expected delay in days from infection to case confirmation and $\mathrm{CAR}(t)$ is the case ascertainment rate on day t. We use a truncated normal prior on the case confirmation delay based on (Jin, 2021):

$$\tau \sim N_{[\tau_D - 8.053 - 1.96 \cdot 4.116, \tau_D]}(\tau_D - 8.053, 4.116),$$

where τ_D is the mean total time from infection to death

$$\tau_D = \delta^{-1} + \gamma^{-1} + \mu^{-1} = 23.6.$$

The underlying mean $m_C(t)$ of $c(t)$ on misreporting days $t = t_0 + k$ is analogous to that for deaths, $m_D(t)$, with the expected number of deaths on accurately reported days t_0, $N\mu R_D(t_0)$, replaced by the expected number of infections confirmed on day t_0, $\tau I_C(t_0)$:

$$\begin{aligned} m_C(t_0) &= \tau I_C(t_0), \\ m_C(t_0 + k) &= \sum_{t=0}^{k-1} m_C(t_0 + k - t)\theta_C^t. \end{aligned}$$

The zero-inflated negative binomial likelihood is then

$$P\left(c(t) = c \mid m_C(t), \kappa_C(t), \theta_C\right) = \begin{cases} \theta_C + (1 - \theta_C) \cdot \mathrm{NegBin2}(0; m_C(t), \kappa_C(t)^{-1}), & c = 0, \\ (1 - \theta_C) \cdot \mathrm{NegBin2}(c; m_C(t), \kappa_C(t)^{-1}), & c > 0, \end{cases}$$

$$\kappa_C(t)^{-1} = \kappa_C^{-1}\left(\zeta_C m_C(t) + (1 - \zeta_C)\right).$$

We use $\mathrm{Uniform}(0, 1)$ priors on θ_C and ζ_C and flat improper priors $\pi(\log \kappa_C) \propto 1$ and $\pi(\log I_C(0)) \propto 1$. We place a beta-distributed random walk prior on case ascertainment rates:

$$\begin{aligned} \mathrm{CAR}(0) &\sim \mathrm{Uniform}(0, 1), \\ \mathrm{CAR}(t + 1) &\sim \mathrm{Beta}(\sigma^2_{\mathrm{CAR}}\mathrm{CAR}(t), \sigma^2_{\mathrm{CAR}}(1 - \mathrm{CAR}(t))), \\ \pi(\log \sigma^2_{\mathrm{CAR}}) &\propto 1. \end{aligned}$$

3.2.3 Effects of NPIs on SARS-CoV-2 transmission

We now turn to the regression model linking NPI policies to the dynamics of viral transmission. Our main source of data is OxCGRT, which aggregates and continuously updates national and subnational government policy responses to the pandemic at the daily level starting from January 1, 2020 (Hale et al., 2021). The database tracks a range of containment and closure, economic, health, and vaccination indicators with numerical values corresponding to the strength of the response on each day. Our transmission regression model focuses on the following 11 policy indicators: school closure, workplace closure, public event cancellation, restrictions on gatherings, public transport closure, stay-at-home requirements, restrictions on internal movement, public information campaigns, testing, contact tracing, and facial covering policies.[4] To account for potential seasonality of SARS-CoV-2 transmission (Gavenčiak et al., 2022; Nichols et al., 2021; Wiemken et al., 2023), we also included average daily temperature measurements reported for the largest population centers in each state, accessed using Meteostat Python (Meteostat, 2022), as a covariate. However, models with temperature as a covariate were excluded based on model selection carried out via leave-one-out cross validation (LOO-CV) (Vehtari et al., 2017).

For each US state s, we define $u^{(s)}(t)$, an 11-dimensional vector with entries in the interval $[0, 1]$ denoting the strength of each NPI implemented on day t. So $u_k^{(s)}(t) = 0$ represents no restrictions associated to the kth NPI on day t (e.g., no school closure), whereas $u_k^{(s)}(t) = 1$ represents the strictest restrictions (e.g., full school closure). NPI implementation during the pandemic was highly correlated, which poses a challenge to teasing apart the effects of individual NPIs on SARS-CoV-2 transmission. We utilize a Bayesian hierarchical model (BHM) to jointly model the time-varying basic reproduction number $R_0^{(s)}(w)$ in each US state

[4]Notably, our model does not include international travel restrictions because: they were a policy held constant in place over time for most of 2020 (hindering identification of their effect on transmission); they were a federal policy (not relevant for state-level decision-making); and they were shown not to be very effective in reducing transmission, only delaying introduction of the virus for a few days (Chinazzi et al., 2020).

s as a function of NPIs using the output from the first stage of estimation described in Section 3.2.2. The BHM leverages spatiotemporal variation in NPI implementation over time across states in order to estimate their effects. It allows for spatial heterogeneity in NPI effects (e.g., due to differential adherence to government mandates) while enabling identification via partial pooling of information across the country.

Propagating uncertainty

Let $\mathcal{D}$ denote the observed case and death data, $\mathcal{U}$ the NPI data, $\mathcal{S}$ the SEIRD model parameters, and θ the NPI regression model parameters. Our two-stage estimation procedure appropriately propagates uncertainty such that the resulting estimates approximate the posterior distribution $\pi(\theta|\mathcal{D},\mathcal{U})$ that would be obtained from combining the epidemiological and regression stages into a single model. Indeed, we have

$$\begin{aligned}
\pi(\theta|\mathcal{D},\mathcal{U}) &= \int \pi(\theta,\mathcal{S}|\mathcal{D},\mathcal{U})d\mathcal{S} \\
&= \int \pi(\theta|\mathcal{S},\mathcal{D},\mathcal{U})\pi(\mathcal{S}|\mathcal{D},\mathcal{U})d\mathcal{S} \\
&= \int \pi(\theta|\mathcal{S},\mathcal{D},\mathcal{U})\pi(\mathcal{S}|\mathcal{D})d\mathcal{S} \\
&\qquad \text{(SEIRD parameters can be inferred from clinical data } \mathcal{D} \text{ alone)} \\
&\approx \frac{1}{M}\sum_{i=1}^{M} \pi(\theta|\mathcal{S}^{(i)},\mathcal{D},\mathcal{U}),
\end{aligned}$$

where $\mathcal{S}^{(i)}$ denotes the ith of M posterior trajectories of the SEIRD parameters (e.g., the time-varying basic reproduction number $R_0^{(s)}(w)$ in each state s) derived from the first-stage transmission model posterior $\pi(\mathcal{S}|\mathcal{D})$. Here we use $M = 100$ randomly sampled trajectories of $R_0^{(s)}(w)$, which define the dependent variable in the NPI regression model. For each $i = 1,\ldots,M$ we generate samples from the NPI model posterior $\pi(\theta|\mathcal{S}^{(i)},\mathcal{D},\mathcal{U})$. We then aggregate these samples to obtain our final estimate of the full posterior $\pi(\theta|\mathcal{D},\mathcal{U})$.

NPI regression model

Our regression model targeting the posterior $\pi(\theta|\mathcal{S},\mathcal{D},\mathcal{U})$ expresses the time-varying basic reproduction number $R_0^{(s)}(w)$ in each state s and week w as a log-linear function of the NPIs implemented in that week, $u^{(s)}(w)$, and the expected number of deaths occurring in the prior week,

$$\text{deaths}^{(s)}(w-1) := \sum_{t:w(t)=w-1} m_D^{(s)}(t),$$

where $m_D^{(s)}(t)$ is as defined in Section 3.2.2 and we obtain weekly values for the NPIs by averaging over days. For notational convenience, we first define the linear predictor

$$\log \hat{R}_0^{(s)}(w) := \log R_0^{(s)} + \beta_u^{(s)} \cdot u^{(s)}(w) + \beta_d^{(s)} \text{deaths}^{(s)}(w-1),$$

where $R_0^{(s)}$ is the initial state-specific basic reproduction number under no restrictions, $\beta_u^{(s)}$ is a vector of state-specific random NPI effects of size $p = 11$, and $\beta_d^{(s)}$ denotes the random effect of deaths. Our final model on the observed transmission rate $R_0^{(s)}(w)$, which is output by the SEIRD model, also includes an autoregressive component:

$$\log R_0^{(s)}(w) = \log \hat{R}_0^{(s)}(w) + \varphi\left(\log R_0^{(s)}(w-1) - \log \hat{R}_0^{(s)}(w-1)\right) + \varepsilon^{(s)}(w), \tag{3.3}$$

where $\varepsilon^{(s)}(w) \sim N(0,\sigma_\varepsilon^2)$ is a normally distributed error term and φ is the AR(1) parameter.[5] The error terms account for unpredictable exogenous shocks that may have sustained effects on transmission (as modeled through the AR(1) term), such as the start of a new wave due to a super-spreader event or the introduction of new infections from an external source (e.g., due to travel into the state).

We control for deaths incident in the prior week following the identification strategy of a number of other studies estimating the causal effects of NPIs (Chernozhukov et al., 2021; Coibion et al., 2020; Crucini & O'Flaherty, 2020; Deb et al., 2022; Goolsbee & Syverson,

[5]We considered AR(q) models for $q = 1, 2, 3$. We selected $q = 1$ via LOO-CV.

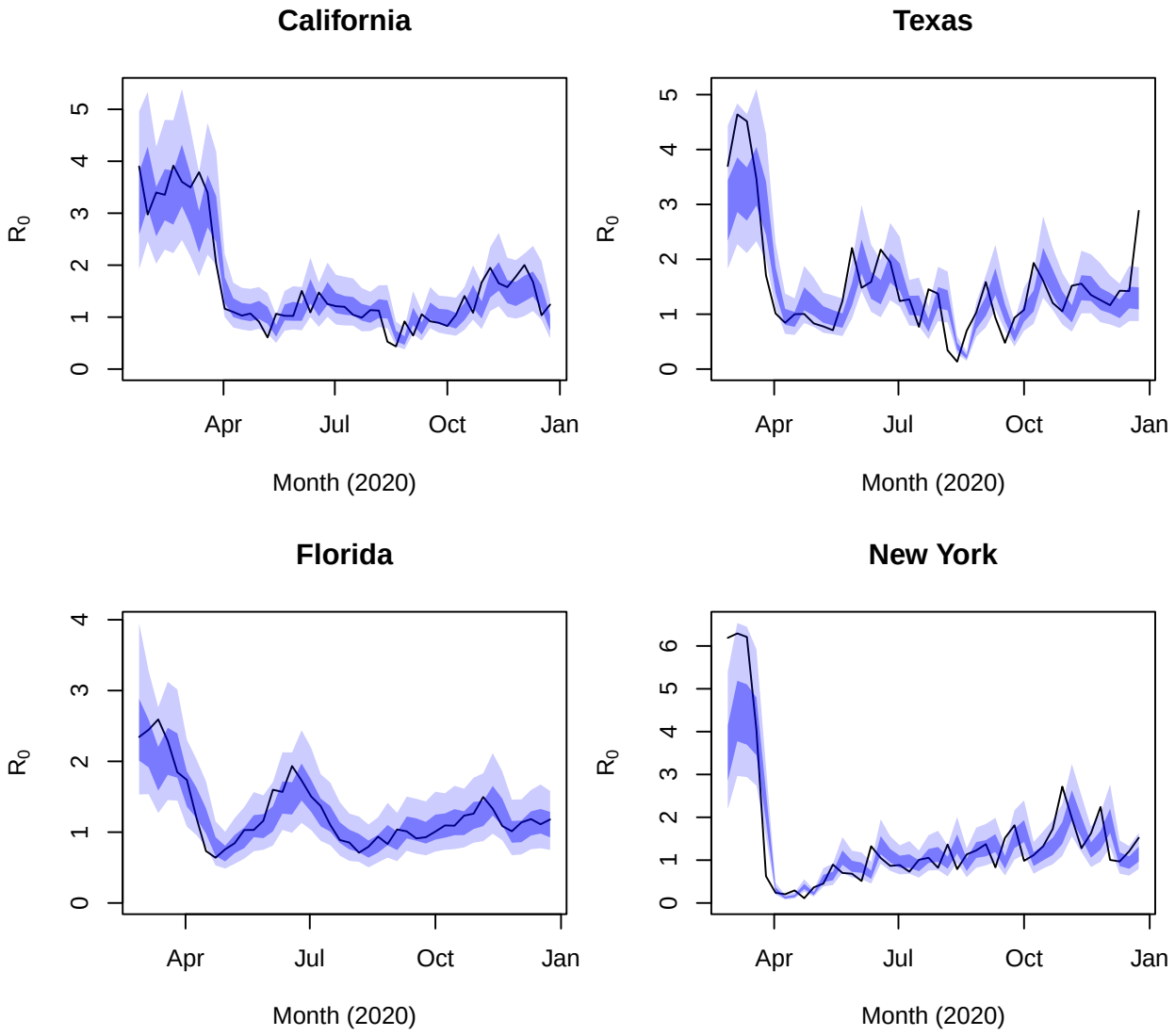


Figure 3.3: Time-varying transmission rates in four states. In black, the MAP trajectory output by the epidemiological model. In dark and light blue, respectively, the 50% and 90% credible intervals of the posterior predictive distribution from the NPI model fitted to this trajectory.

2021; Karaivanov et al., 2021; Mader & Rüttenauer, 2022; Verschuur et al., 2021). The existence of substantial voluntary social distancing and its pronounced economic effects in the US and elsewhere have been well-documented in numerous empirical analyses (Abouk & Heydari, 2021; Aum et al., 2021; Badr et al., 2020; Bartik et al., 2020; Bodenstein et al., 2022; Chen et al., 2020; Chernozhukov et al., 2021; Cronin & Evans, 2020; Goolsbee & Syverson, 2021; Jamison et al., 2021) and derived from first principles in macroeconomic modeling (Brzezinski et al., 2020; Eichenbaum et al., 2021; Farboodi et al., 2021; Krueger et al., 2022). In response to SARS-CoV-2 outbreaks, people began social distancing (and, potentially, other protective measures, e.g., mask-wearing) prior to the onset of restrictions and subsequently increased social activity separate from the lifting of restrictions. As a result, declines in mobility, consumer spending, and hours worked cannot be fully attributed to the effects of NPI policies (Baker et al., 2020; Forsythe et al., 2020). Including prior deaths as a covariate in the regression model accounts for changes in protective behaviors by individuals responding to the risk of infection.[6]

Figure 3.1b depicts a directed acyclic graph (DAG) representing our causal model for the outcome $R_0(w)$ in each week. We suppress the state s for compactness of notation. Deaths incident in week $w-1$ may affect the policy response $u(w)$ and the transmission rate $R_0(w)$ in the following week, with the latter effect representing the endogenous behavioral response (including social distancing and other protective measures) to the fear of infection. The transmission rate $R_0(w)$ is also a function of NPI policies $u(w)$ and exogenous shocks $\varepsilon(1:w) := \{\varepsilon(v) : v = 1, \ldots, w\}$.[7] We allow for past shocks $\varepsilon(1:w-1)$ to affect past

[6]This model was selected by LOO-CV among models controlling for both deaths and cases in the past x weeks, where x was fixed at $1, 2, 4, 6, 8, 10, 12, 16$ and also allowed to vary as a parameter in the model.

[7]In line with a number of other studies estimating the effects of NPIs (e.g., Brauner et al. (2021), Flaxman et al. (2020), and Sharma et al. (2021)), we do not account for lagged effects of NPIs. This is not an issue for closure policies (e.g., school, business, and transit closures), as they take effect immediately. For other policies, it is reasonable to assume that behavioral responses occur quickly. For example, Alexander and Karger (2023) find that mobility and consumer spending declined consistently within two days of when US counties enacted stay-at-home orders.

deaths and the current policy response $u(w)$. In our regression model (3.3), the lagged and attenuating effect of past shocks on the transmission rate is captured by the AR(1) term:

$$\varphi\left(\log R_0^{(s)}(w-1) - \log \hat{R}_0^{(s)}(w-1)\right) = \sum_{v=1}^{w-1} \varphi^{w-v} \varepsilon^{(s)}(v).$$

Given the DAG 3.1b, we see that controlling for deaths$(w-1)$ and $\varepsilon(1:w-1)$—as we do in (3.3)—blocks all back-door paths from $u(w)$ to $R_0(w)$. As such, the effect of NPIs is identified following the back-door criterion (Pearl, 2009).

Regarding prior specification for the regression (3.3), we use a hierarchical model for the state-specific coefficients $\theta^{(s)} = (\beta_u^{(s)}, \beta_d^{(s)}, R_0^{(s)}) \in \mathbb{R}^{p+2}$, which enables partial pooling of information. With θ denoting the global pooled effects, we have

$$\begin{aligned}
\theta^{(s)} &\sim \text{Normal}(\theta, V) \prod_{k=1}^{p} I\left(\theta^{(s)}(k) \leq 0\right), \qquad (3.4)\\
V &= D(\lambda)\Omega D(\lambda),\\
\Omega &\sim \text{LKJ}(\zeta = 1),\\
\lambda(j) &\sim \text{Student-}t_{[0,\infty)}(0, 2.5^2, 3), \quad j = 1, \ldots, p+2,\\
\sigma_\varepsilon &\sim \text{Student-}t_{[0,\infty)}(0, 2.5^2, 3),\\
\varphi &\sim \text{Uniform}(-1, 1).
\end{aligned}$$

The truncated normal prior (3.4) on the state-level random effects $\theta^{(s)}$ assumes that they are centered around the pooled effects θ and that NPIs cannot increase the transmission rate ($\theta^{(s)}(k) \leq 0$), in line with the results of numerous studies estimating the effects of NPIs (Banholzer et al., 2021; Brauner et al., 2021; Flaxman et al., 2020; Haug et al., 2020; Hsiang et al., 2020; Jamison et al., 2021; Li et al., 2021; Liu et al., 2021; Sharma et al., 2021; Stokes et al., 2022). For the remaining parameters, we use the default prior specification for multilevel models used in the brms R package (Bürkner, 2017). The covariance matrix V of the random effects is decomposed as the product of a correlation matrix Ω given an

LKJ prior with parameter $\zeta = 1$ (specifying a uniform prior on correlation matrices), and a diagonal matrix $D(\lambda)$ with entries $\lambda(j)$ given Student-t priors with 3 degrees of freedom truncated to be non-negative. The error standard deviation σ_ε is given the same truncated Student-t prior. The AR(1) parameter φ is given a Uniform$(-1, 1)$ prior.

3.2.4 Evaluating and optimizing costs

The SEIRD equations (3.1) combined with the NPI regression model (3.3) define a simulator for the trajectories of infections and deaths under counterfactual NPI policies, conditional on the parameters estimated using the NPI and clinical data—denoted $\mathcal{U}$ and $\mathcal{D}$, respectively. Evaluating the cost-effectiveness of NPIs and determining optimal strategies requires accounting for the aggregate costs incurred in implementing policies and their consequent health impacts. For an NPI policy $\mathbf{u} = \{u(t)\}_{t=1}^{T}$ implemented in a state on the days $t = 1, \ldots, T$, we define its associated cost $\mathcal{C}(\mathbf{u})$ as the sum of the posterior expected costs incurred by infections and NPI implementation:

$$\mathcal{C}(\mathbf{u}) = \mathbb{E}\left[c_{\text{NPI}}(\mathbf{u}) + \frac{1}{N}\sum_{t=1}^{T} c_\nu \nu(t) \middle| \mathcal{D}, \mathcal{U}\right], \tag{3.5}$$

where $\nu(t) = N\beta(t)S(t)I(t)$ is the number of new COVID infections in the state incident on day t under the policy $\mathbf{u}$. Here c_ν is the average cost in USD2020 associated to a COVID infection and

$$c_{\text{NPI}}(\mathbf{u}) = \sum_{k=1}^{p} c_k(\mathbf{u}_k)$$

is the average per capita cost in USD2020 associated to implementing the policy $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$, where c_k is the cost of the kth NPI. We define these quantities below.[8] Note that, in practice, we cannot directly evaluate the expectation (3.5). Instead, we use posterior trajectories of infections $\nu(t)$ to approximate (3.5) via Monte Carlo.

[8]Given the relatively short duration of our study period (i.e., the first year of the pandemic), we do not adjust the cost function for temporal discounting.

Cost of infections

The average cost of a COVID infection is a sum of average life costs (due to COVID deaths), medical costs (incurred by treatment), productivity costs (due to worker absenteeism), and costs associated to voluntary social distancing in response to the fear of infection.

The life cost associated to a COVID infection deserves some discussion, as it dominates the aggregate cost of infection and it requires ascribing a dollar value to death, which can be a contentious issue. In cost-benefit analysis, the standard approach to quantifying mortality risk reductions as a result of public policy in monetary terms is through the value of a statistical life (VSL), commonly estimated at about $11 million in USD2020 (Greenstone & Nigam, 2020). In our context, the relevant quantity is the value of a statistical COVID death (VSCD). As Robinson et al. (2021) note, COVID deaths are concentrated in the oldest age groups and the decision to adjust the VSCD for the age profile of COVID mortality or not can vastly alter the conclusions of a cost-benefit analysis. As such, we conduct sensitivity analysis of our results using high and low estimates of the VSCD reported in (Robinson et al., 2021). In line with a number of other studies in the COVID economics literature (Farboodi et al., 2021; Greenstone & Nigam, 2020; Hall et al., 2020; Kaplan et al., 2022), we use as our baseline the low estimate of $4.47 million, which is based on the average years of life lost to a COVID death and a constant value per statistical life year.[9] The high estimate of $10.63 million assumes the VSCD equals the population average VSL (i.e., it does not adjust the VSL for the age pattern of COVID deaths). Tables 3.2 and 3.3 record the value of this and other economic parameters used in our study. Denoting the VSCD by c_{VSCD}, the average life cost per COVID infection in state s is then $c_{\text{VSCD}} \cdot \iota^{(s)}$, where $\iota^{(s)}$ is the state-level IFR. Using the baseline VSCD, this quantity ranges from $10,000 to $76,000 across states with a median of $34,000.

[9]That is, outside of our sensitivity analysis (Section 3.3.3), all estimated costs are based on the low VSCD.

To quantify the cost associated to voluntary social distancing, we rely on the results of Aum et al. (2021), who find that a one per thousand increase in the COVID case rate caused a 2.68% drop in employment in South Korea in the spring of 2020, with similar (although non-causal) estimates in the US and UK. We adjust their numbers for under-reporting of cases based on the number of deaths and cases in South Korea in their period of study reported by Our World in Data (Mathieu et al., 2020). By February 29, 2020, South Korea had 556 cumulative confirmed cases, and by March 7 it had 3,526. The cumulative COVID deaths three weeks later on March 21 were 75, and by March 28 were 104. With an IFR of 0.68% (Meyerowitz-Katz & Merone, 2020), we would expect between 11,029 and 15,294 infections with this number of deaths. This suggests a case ascertainment rate between 5% and 23% in that period. We average these two numbers, assuming 15.5% case ascertainment, which yields a $0.155 \cdot 2.68 \approx 0.42\%$ drop in employment resulting from a one per thousand increase in COVID prevalence. In line with Chetty et al. (2024), Aum et al. (2021) find that these impacts were felt most acutely among low-wage workers. As such, we phrase this increase in unemployment due to an infection—which is assumed to last the average duration of the infectious period ($\gamma^{-1} = 8.5$ days)—in monetary terms using the state's median income.[10] In sum, we find that the fear cost per COVID infection ranges from \$1,351 to \$7,241 across states with a median of \$2,133.

For the remaining parameters, we take \$3,045 as the average medical cost of a COVID infection based on the estimates of Bartsch et al. (2020) and DeMartino et al. (2022). We assume the average productivity cost of a COVID infection is equal to one week of sick days at a state's median wage, which yields costs similar to those reported by Skarp et al. (2021). These range from \$505 to \$1,084 across states with a median of \$677.

[10]We use national and state-level personal income data reported by U.S. Bureau of Economic Analysis (2023b). To approximate state-level median income, we multiply the US median personal income by the state's per capita income divided by the US per capita income.

Cost of workplace closures and social distancing measures

Beginning in the spring of 2020, consumer spending dropped significantly in response to health concerns and government-mandated business closures and social distancing policies. This reduction in consumer spending—primarily on in-person services—was responsible for a large majority of the decline in US GDP in the second quarter of 2020. Declining business revenue led to substantial layoffs with subsequent unemployment increases concentrated among low-wage workers (Chetty et al., 2024). As such, we quantify the cost of workplace closures and social distancing measures through their effects on employment, which have been thoroughly studied in the COVID economics literature (Alexander & Karger, 2023; Baek et al., 2021; Barrot et al., 2024; Bartik et al., 2020; Bodenstein et al., 2022; Coibion et al., 2020; Crucini & O'Flaherty, 2020; Gupta et al., 2023). As above, we convert employment rate decreases in each state to monetary losses using the state's median personal income, which reflects the wage distribution of pandemic job loss.

The effects of COVID workplace closures on unemployment and consumer spending in the US were estimated by Barrot et al. (2024), Crucini and O'Flaherty (2020), and Gupta et al. (2023), who arrive at broadly similar conclusions. Crucini and O'Flaherty (2020) found that non-essential business closures led to a 1-2 percentage point decline in expenditures. Gupta et al. (2023) found that 60% of the 12 percentage point decline in the employment rate between January and April 2020 was due to state policies, with government-mandated business closures and stay-at-home orders each accounting for half of those 7.2 percentage points.[11] Barrot et al. (2024) found that a 10 percentage point increase in the share of restricted labor was associated with a 3 percentage point decline in April 2020 employment. Given these findings, we define low, middle, and high scenarios in which workplace closures cause a 2%, 4%, and 6% declines in the employment rate. We take the middle scenario as

[11] Although drops in consumer foot traffic are not directly comparable to employment rate decreases, Goolsbee and Syverson (2021) found that general shelter-in-place orders reduced overall consumer visits by 7 percentage points.

our baseline and consider the low and high scenarios in our sensitivity analysis in Section 3.3.3.

We define social distancing measures as the combination of the following six NPIs tracked by OxCGRT: stay-at-home orders, restrictions on gatherings, restrictions on internal movement, public information campaigns, public transit closures, and public event cancellations. We bundle these policies for a number of reasons: their implementation was highly correlated in time and space; there is a paucity of information on the individual economic effects of most of these interventions as the COVID economics literature tends to focus on "social distancing" or "lockdown" measures broadly defined (likely due to their synchronous adoption); and they are blanket policies acting as relatively blunt instruments with their primary direct effects on the economy stemming from a common mechanism—namely, reduction in consumer spending on in-person services with consequent unemployment.

The effects of social distancing measures on unemployment and consumer spending in the US were studied by Baek et al. (2021), Bodenstein et al. (2022), Coibion et al. (2020), Crucini and O'Flaherty (2020), and Gupta et al. (2023). Drawing on survey responses, Coibion et al. (2020) found that individuals in counties under lockdown were 2.8 percentage points less likely to be employed relative to other survey participants, had a 1.9 percentage point lower labor-force participation, and had a 2.4 percentage point higher unemployment rate. Crucini and O'Flaherty (2020) found that stay-at-home orders caused a 4 percentage point decrease in consumer spending and hours worked. Bodenstein et al. (2022) found that the combined effect of voluntary and mandatory social distancing could explain 6–8 percentage points of the 12% drop in US GDP in the second quarter of 2020 and that stay-at-home orders could account for a 2 percentage point increase in the unemployment rate. As mentioned above, Gupta et al. (2023) found that stay-at-home orders led to a 3.6 percentage point decline in employment rates through April 2020. Similarly, Baek et al. (2021) found that each week of stay-at-home order exposure between March 14 and April 4, 2020 yielded an increase in

a state's weekly unemployment insurance claims corresponding to 1.9% of its employment level. As Bartik et al. (2020) note, nearly all employment declines occurred within the two-week period March 14–28, which implies a cumulative 3.8% drop in the employment rate based on the findings of Baek et al. (2021). Hence, we assume that social distancing measures cause a 4% decline in the employment rate. In our sensitivity analysis, we do not vary the cost of social distancing measures as we are primarily interested in assessing the robustness of the optimal strategy and the relative costs of various policies rather than variation in the total cost incurred by each policy, which means that we are free to leave the value of one term in the cost function (3.5) fixed.

We note that the combined economic effects of workplace closures and social distancing measures used here are on par with trends in aggregate economic output in the US and elsewhere observed in 2020 and in prior pandemics. Congressional Budget Office (2020) estimates a 3.5% year-over-year decline in real US GDP from 2019 to 2020. Analyzing trends in annual global GDP, Kaplan et al. (2022) estimate a 7% decline from 2019 to 2020 due to COVID, which equates to a loss of $10 trillion. Demirgüç-Kunt et al. (2021) find that national lockdowns led to a 10% decline in economic activity across Europe and Central Asia in the spring of 2020. Studying the 1918 Spanish flu pandemic, in which social distancing measures were the primary tools used to curtail viral spread, Barro et al. (2020) estimate a cumulative loss in GDP per capita of 6% over 3 years.

Cost of school closures

The cost of school closures is a sum of productivity loss due to worker absenteeism (as parents of children out of school miss work to care for their kids) and learning loss resulting from students missing school and receiving lower quality education through distance learning.

Lempel et al. (2009) and Sadique et al. (2008) estimated the magnitude of direct GDP loss due to worker absenteeism resulting from extended school closures in the US and UK,

respectively, and arrived at nearly identical numbers. They find that four weeks of school closure would cost 0.1–0.3% of GDP in the US and 0.1–0.4% in the UK. For our study, we use 0.2%. Similar estimates based on modeling studies are reviewed in Viner et al. (2020).

Notably, Lempel et al. (2009) also estimate the healthcare impacts of a four-week school closure in the US, finding that it would lead to a reduction of 6% to 19% in key healthcare personnel. Similarly, Bayham and Fenichel (2020) find that 15% of the healthcare workforce would be in need of childcare during a school closure and find that their absence from work could cause a greater number of COVID deaths than school closures prevent. Pricing these health impacts is not straightforward, so we omit these considerations when defining the cost function. As such, we believe that our accounting of the costs of school-closure-related worker absenteeism is conservative.

While learning loss due to school closure can be viewed as a social cost, it can lead to substantial downstream economic costs as cohorts of students that missed significant schooling eventually enter the labor-force as less skilled and productive workers. Education economists have extensively studied the connections between time spent in school, performance on standardized tests, and subsequent impacts on lifetime earning and GDP with findings that are consistent across contexts. Hanushek and Woessmann (2020) and Psacharopoulos et al. (2021)—whose assessments of the cost of learning loss we use—provide discussion and references. As our high scenario, we use the estimate of Hanushek and Woessmann (2020), who find that cohort learning loss equivalent to one-third of a school-year has a staggering net present value equal to 69% of current-year GDP.[12] Psacharopoulos et al. (2021) arrive at a much smaller number, finding a 9% GDP loss arising from 0.33 years of lost schooling, which

[12]Hanushek and Woessmann (2020) also estimate that a student missing 0.33 years of school leads to a loss in lifetime individual income of 3.0% in the US and 2.6% pooled globally. Fuchs-Schündeln et al. (2022) find average losses of 2.1% in lifetime earnings and 1.2% in permanent consumption of children affected by COVID-19 school closure. Considering that Betthäuser et al. (2023) report a learning deficit of 0.35 school-years accrued during COVID, the estimates of Hanushek and Woessmann (2020) and Fuchs-Schündeln et al. (2022) are quite similar.

forms our low scenario and also our baseline value. We note that the results of Psacharopoulos et al. (2021) are predicated on the assumption that remote learning is 90% as effective as in-person school, which is a likely source of the large discrepancy between the two estimates. While distance learning certainly mitigated some learning loss (Betthäuser et al., 2023), and keeping schools open during the pandemic would have also incurred some learning loss due to student and teacher illness-related absenteeism, we believe that this assumption leads to a conservative estimate of the cost of learning loss associated to in-person school closure.[13] Nevertheless, as we discuss in Section 3.3, we find that optimal NPI strategies based on this low estimate involve no closure of schools beyond the usual 16 weeks of break per year.

In their systematic review and meta-analysis, Betthäuser et al. (2023) find a substantial and consistent learning deficit of 0.35 school-years of learning loss across 15 high- and middle-income countries, which accrued early in the pandemic.[14] This learning gap persisted but ceased to grow beyond 0.35 school-years, which suggests that remote learning did mitigate learning loss with greater efficacy (relative to in-person schooling) as time went on.[15] In our cost function, we account for the improving quality of remote learning over time by assuming that the amount of learning loss incurred by one week of school closure equates to one school-week initially and decreases linearly to 0 as a function of the cumulative number of past weeks spent under school closure, such that 0.35 school-years is the maximal cumulative amount of learning loss possible. Furthermore, our cost function only accounts for marginal learning loss (i.e., beyond what would be expected after summer break, for example) by

[13]Indeed, based on the results of a recent preprint (Fahle et al., 2023), Mervosh et al. (2024) demonstrate that drops in math scores in mostly in-person school districts were only 2/3 of those in mostly remote or hybrid districts among third through eighth graders in the U.S. during COVID-19.

[14]Citing (Fahle et al., 2023), Mervosh et al. (2024) note that aggregate learning loss in the U.S. during COVID-19 likely exceeded 0.35 school-years. Again, we believe that our accounting of the costs of school closure is conservative as such.

[15]Similarly, based on results of a simulation model published earlier in the pandemic, Azevedo et al. (2021) projected that COVID-19 school closures could result in learning loss equivalent to 0.3–1.1 years of schooling.

assuming that learning loss only begins to accrue once the duration of school closure exceeds 16 weeks.

Cost of testing, tracing, and masking

We quantify the per capita cost of a week-long mask mandate as the price of supplying an individual with masks for a week. Following Bartsch et al. (2022), we assume personal mask expenditure of \$0.32 per day or, equivalently, \$2.24 per week, which approximates the cost of one surgical mask per day or one N95 mask per week (Skarp et al., 2021).

In the first year of the pandemic, US states steadily ramped up the number of SARS-CoV-2 PCR tests administered each day at a consistent linear pace. Indeed, after running least-squares regression of the cumulative number of tests administered in a state on each day against time (squared) using test data obtained from the COVID Tracking Project (The Atlantic, 2021), we obtain R^2 values above 0.97 for all states. Across states, the linear rate of testing capacity increase varies from an additional 7 to 40 tests per million population per day. Our cost function accounts for this by assuming that the number of tests administered in a given week under mask mandate is a linear function of the cumulative number of past weeks spent under mask mandate, with the slope given by the state-specific rate of testing capacity increase obtained from the regression. This yields the total number of tests administered in a state in any given week under the specified masking policy. We convert this quantity to a dollar value assuming that each test costs \$100 based on Lo et al. (2023), Sharfstein (2024), and Skarp et al. (2021), which includes the cost of procuring the test as well as labor for sample extraction and diagnostic lab testing.

We similarly assume that, while contact tracing policies are in place, tracing capacity ramps up at a linear pace. This is in line with increases over time in capacity reported in wide-scale assessments of US contact tracing programs (Lash et al., 2021; Rainisch et al., 2022), as well as general increases over time in state-level hiring of contact tracers reported

in media (Henry, 2020; Simmons-Duffin, 2020). Fitting a log-normal model to data from Lash et al. (2021), we estimate the mean number of cases interviewed per week per 100,000 population to be 95.0 during their period of study, June–October 2020. Similarly, based on data from Rainisch et al. (2022), we estimate a mean of 170.5 cases interviewed per week per 100,000 population during November 2020–January 2021. With four months separating August and December 2020 (the midpoints of the respective study periods), we therefore assume an increase in capacity of $(170.5 - 95.0)/16 \approx 4.72$ cases interviewed per 100,000 population per week while contact tracing policies are active. We convert this number to a dollar value based on the average cost of contact tracing per index case. Fields et al. (2021) report the hourly cost of contact tracing at $\$107.22/4.16 \approx \$25.77$. According to Spencer (2021), the median caseload per investigator during their two-week evaluation period was 31. Assuming a 40-hour work week, this implies a cost per case of $\$25.77 \times 80/31 \approx \$66.50$. This number, which we take as our cost of contact tracing per index case, is near the midpoint of the interval reported in Skarp et al. (2021) ($40.73–$93.59) based on different data. Table 3.4 records the testing, tracing, and masking cost parameters with references.

3.3 Results

3.3.1 Epidemiological model

Figure 3.4 displays estimates of active viral prevalence and posterior predictive distributions of the observed deaths and cases in the US in 2020. While the posterior predictive distributions of deaths are well-calibrated at the state-level (as evidenced by Figure 3.2, for example), when aggregated to the US as a whole they exhibit under-coverage. This is because we model the states independently and do not explicitly account for the "weekend effect", i.e., consistent under-reporting of deaths on weekends which leads to highly correlated residuals across states on those days. As we evaluate and optimize NPI policies at the state-level, however, this does not pose an issue for our downstream analysis. See Irons and Raftery (2021) for

Table 3.2: Economic parameters. All costs are in USD2020.

Parameter	**Value**	**Reference**
2019 GDP per capita by state	Varying ($40,600–219,000)	U.S. Bureau of Economic Analysis (2023a)
2019 per capita income by state	Varying ($39,400–84,700)	U.S. Bureau of Economic Analysis (2023b)
2019 US median personal income	$35,980	U.S. Census Bureau (2023)
2019 population by state	Varying (0.575–39.5 million)	U.S. Bureau of Economic Analysis (2022)
2019 US GDP current dollar growth rate	4.1%	U.S. Bureau of Economic Analysis (2020)
Value of a statistical COVID death (VSCD)	Low: $4.47 million High: $10.63 million	Robinson et al. (2021)
Voluntary social distancing cost per COVID infection by state	Varying ($1,351–7,241)	Aum et al. (2021)
Productivity cost of COVID infection	One week of state median income	Skarp et al. (2021) and U.S. Bureau of Economic Analysis (2023b)

Table 3.3: Economic parameters (continued). All costs are in USD2020.

Parameter	**Value**	**Reference**
Average medical cost of COVID infection	$3,045	Bartsch et al. (2020) and DeMartino et al. (2022)
Net present value of GDP loss due to learning loss by state	Low: 9% GDP per 0.33 school-years High: 69% GDP per 0.33 school-years	Psacharopoulos et al. (2021) Hanushek and Woessmann (2020)
Learning loss accrued during COVID	0.35 school-years	Betthäuser et al. (2023)
Direct GDP loss due to school closure	0.2% GDP per four weeks	Lempel et al. (2009)
Employment rate decrease due to workplace closure	Low: 2%; Mid: 4%; High: 6%	Barrot et al. (2024), Crucini and O'Flaherty (2020), and Gupta et al. (2023)
Employment rate decrease due to social distancing mandates	4%	Baek et al. (2021), Bodenstein et al. (2022), Coibion et al. (2020), Crucini and O'Flaherty (2020), and Gupta et al. (2023)

Table 3.4: Masking, testing, and tracing parameters. All costs are in USD2020.

Parameter	**Value**	**Reference**
Daily personal mask expenditure	$0.32	Bartsch et al. (2022) and Skarp et al. (2021)
Cost of a PCR test	$100	Lo et al. (2023), Sharfstein (2024), and Skarp et al. (2021)
Daily rate of testing capacity increase by state	Varying (7–40 tests per million pop.)	The Atlantic (2021)
Cost of contact tracing per index case	$66.50	Fields et al. (2021), Skarp et al. (2021), and Spencer (2021)
Weekly rate of contact tracing capacity increase	4.72 cases per 100k pop.	Lash et al. (2021) and Rainisch et al. (2022)

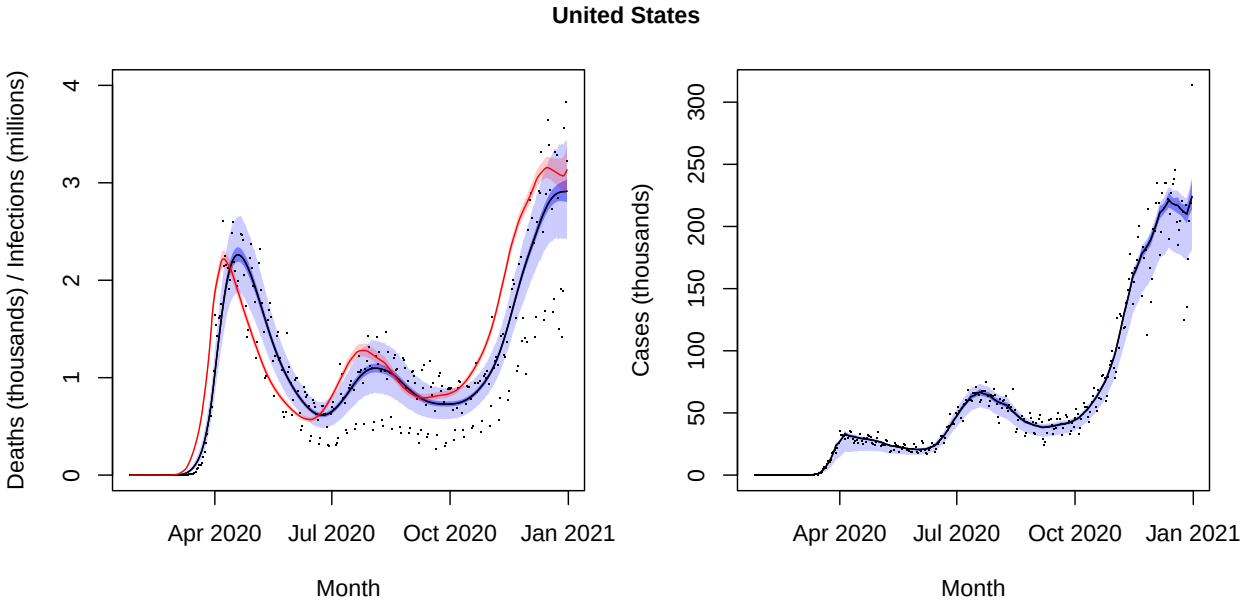


Figure 3.4: State-specific SEIRD results aggregated to the US. Observed deaths $d(t)$ and cases $c(t)$ are plotted in black. Posterior median and 90% credible intervals of the underlying mean parameters $m_D(t)$ and $m_C(t)$ are in dark blue. 90% credible intervals of the posterior predictive distributions of $d(t)$ and $c(t)$ are in light blue. On the left, posterior median and 90% credible intervals for active viral prevalence $I(t)$ are in red.

more detailed reports and discussion of state-specific prevalence estimates.

We estimate that there were 52.9 (95% CI: 51.9–54.1) million COVID infections in the US in 2020, representing about 16% of the population. As there were about 397,000 COVID deaths in the US through January 22, 2021 (Mathieu et al., 2020) (accounting for the three weeks, on average, between infection and death), this corresponds to an IFR of about 0.75%. Our findings are in line with the systematic meta-analysis of Meyerowitz-Katz and Merone (2020), who estimated an IFR of 0.68% (0.53%–0.82%) for COVID in 2020 based on 24 studies from a range of countries, as well as Eales et al. (2023), who estimated the IFR

in England in 2020 based on a series of nationally representative testing surveys at 0.67% (0.65%–0.70%). Similarly, Ward et al. (2024) estimated the IFR in England in October 2020 at 0.74% (0.48%–1.40%).

Our estimates of SARS-CoV-2 infections incident in 2020 leverage prior work based on random sample testing (Irons & Raftery, 2021)—a putatively unbiased measure of viral prevalence—and, as noted above, produce an IFR similar to that estimated in England in 2020 also based on representative testing surveys. Nevertheless, we note that our findings concerning policy evaluation and optimization below are robust to sensible variations in the IFR. This is because the costs of infections are dominated by COVID deaths, which are identified from the clinical data we use here and, therefore, are outputs of our model not substantially affected by the IFR parameter (which only varies the estimated number of infections incident per death).

3.3.2 NPI regression model

Fit to data

Regarding the fit to data, the posterior median R^2 for the log-linear regression model, defined in terms of the empirical variance of the transmission rates $\widehat{\text{Var}}(\log R_0^{(s)}(w))$ as

$$R^2 = 1 - \frac{\sigma_\varepsilon^2}{\widehat{\text{Var}}(\log R_0^{(s)}(w))},$$

is 0.71, indicating that a substantial proportion of the variance in transmission rates remains unexplained by NPIs or behavioral response to the fear of infections. Indeed, there are numerous factors affecting SARS-CoV-2 transmission—including super-spreader events and introduction of new infections from outside the state—accounted for by the error terms $\varepsilon^{(s)}(w)$ that are difficult to predict. Figure 3.3 plots the *maximum a posteriori* (MAP) trajectory of R_0 in the four most populous states—California, Texas, Florida, and New York—against the posterior predictive distribution from the NPI model fit to this trajectory.

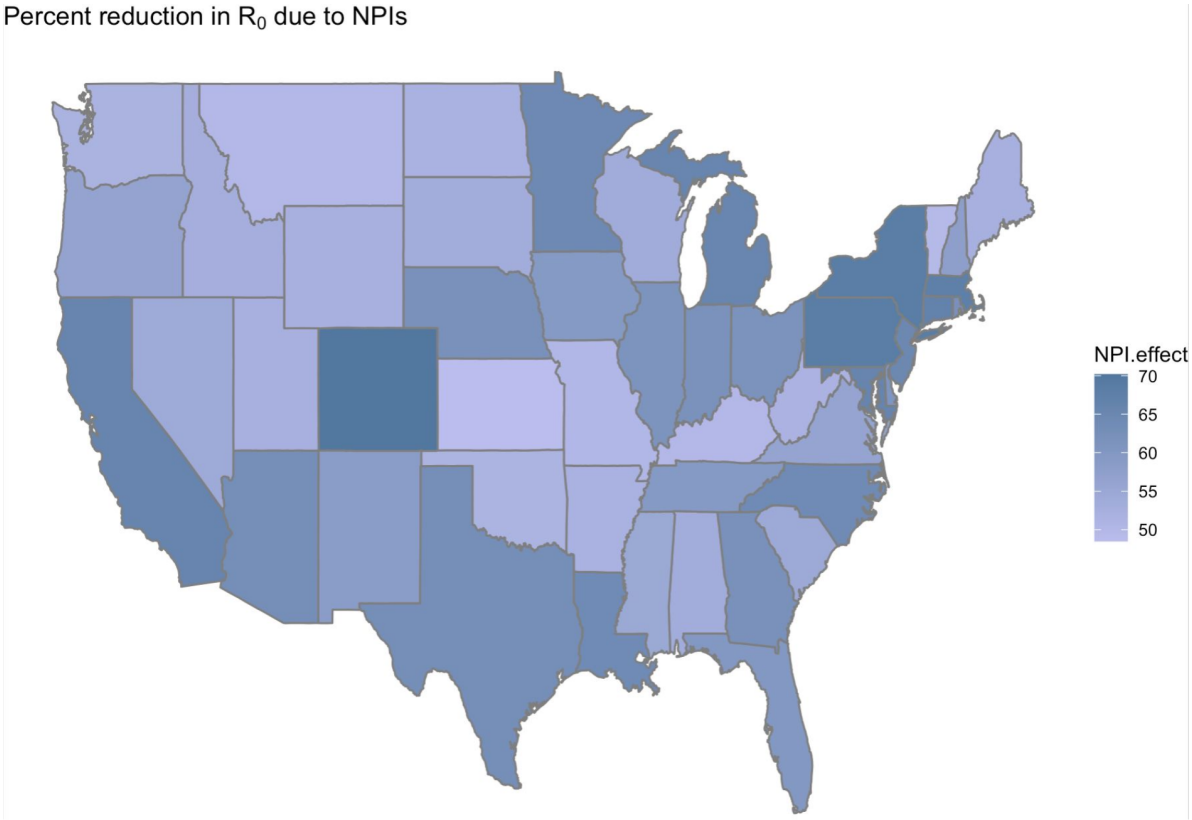


Figure 3.5: Posterior median total percent reduction in R_0 due to NPIs by state.

The model fits the data well, but cannot capture unpredictable shocks in transmission, which are reflected in future predicted values of the transmission rate through the AR(1) term. The AR(1) parameter φ is 0.77 (0.74–0.81), indicating a high degree of residual autocorrelation in the time-varying reproduction number $R_0^{(s)}(t)$ across states.

Total effect of NPIs

We estimate a pooled baseline reproduction number R_0 under no interventions of 2.5 (95% CI: 2.2–2.9). This comports with the systematic review of Liu et al. (2020), who report a median R_0 of 2.79 for wild-type SARS-CoV-2. The pooled total effect of NPIs, $\alpha = \sum_{k=1}^{p} \beta_u(k)$,

which represents the effect of "full lockdown"—i.e., $u_k(w) = 1$ for all $k = 1, \ldots, p$—yields a reduction of R_0 by 58.3% (52.2%–64.4%) to 1.05 (0.96–1.15). By comparison, the pooled effect of deaths, β_d, is -0.55 (-0.76– -0.35), which translates to a 5.3% (3.4%–7.3%) decrease in R_0 when the death rate in the prior week increases by 1 per 100,000 population. To illustrate, we note that the mean weekly death rate across states in 2020 was 2.7 per 100,000 population, which yields a 13.7% (8.9%–18.5%) reduction in R_0—between about one fifth and one third of the total effect of NPIs—from voluntary social distancing and other protective measures due to fear of infection.

For a population of this kind, NPIs alone would most likely not be sufficient to suppress transmission at the start of the outbreak ($R_0 < 1$). However, voluntary protective measures and acquired immunity in combination with full lockdown would be enough to extinguish viral spread (at least in the absence of exogenous shocks). Figure 3.8 exhibits posterior trajectories of deaths in the US under various NPI strategies. The posterior median cumulative deaths in 2020 under full lockdown would have been 73,427, about one fifth of the 348,949 actually observed.

Our estimate of the total percent reduction in R_0 due to NPIs is more conservative than others reported in the literature. Flaxman et al. (2020), Brauner et al. (2021), and Banholzer et al. (2021), respectively, find 81% (75%–87%), 77% (67%–85%), and 67% (64%–71%) reductions in transmission in the initial spring 2020 wave. Studying the second wave, Sharma et al. (2021) report a combined NPI effect of 66% (61%–69%). We note that none of these studies control for confounding (e.g., endogenous social distancing), which may account for the discrepancy with our estimates. Indeed, when we add the effect of deaths to that of NPIs, the combined reduction in R_0 approaches these higher estimates. Another possible explanation is the context: we study the US whereas Banholzer et al. (2021), Brauner et al. (2021), Flaxman et al. (2020), and Sharma et al. (2021) focus primarily on European countries, which may have implemented stricter NPIs or practiced greater adherence to restrictions,

and which exhibited higher R_0 values (3.3–3.8), perhaps due to earlier introduction of the virus to European countries or higher levels of social mixing on average.

Zooming in on the state-level results, we can similarly quantify the total effect of NPIs on transmission in state s by

$$\alpha^{(s)} := \sum_{k=1}^{p} \beta_u^{(s)}(k),$$

with $p^{(s)} = 100(1 - \exp(\alpha^{(s)}))$ representing the total percent reduction in $R_0^{(s)}$ under full lockdown. Figure 3.5 displays the geographic distribution of the posterior median of $p^{(s)}$ across states, which ranges from 49% to 70%. Overall, NPIs tend to be more effective in more urbanized and populous states. Some of this variation may be explained by the literature on political polarization and partisan social distancing during the pandemic (Adolph, Amano, Bang-Jensen, Fullman, Magistro, Reinke, Castellano, et al., 2022; Adolph, Amano, Bang-Jensen, Fullman, Magistro, Reinke, & Wilkerson, 2022; Adolph et al., 2021; Allcott et al., 2020; Barrios & Hochberg, 2021; Brodeur et al., 2021; Painter & Qiu, 2021). Alternatively, and related to our discussion in the previous paragraph, we note that rural states tend to have lower baseline $R_0^{(s)}$ values, possibly due to later importation of the virus and lower levels of social mixing. With a lower ceiling in these states, NPIs have less room to suppress transmission.

Finally, note that the p-vector

$$\rho^{(s)} := \beta_u^{(s)} / \alpha^{(s)}$$

consists of weights representing the proportional contribution of each NPI to the total reduction of transmission. Here $\rho^{(s)}$ can be thought of as defining a data-driven "stringency index" combining NPIs according to their strengths in a single-number summary of the stringency of government restrictions, as opposed to previously defined measures, such as OxCGRT's stringency index, which average NPIs uniformly without regard for their varying effects on transmission (Hale et al., 2021).

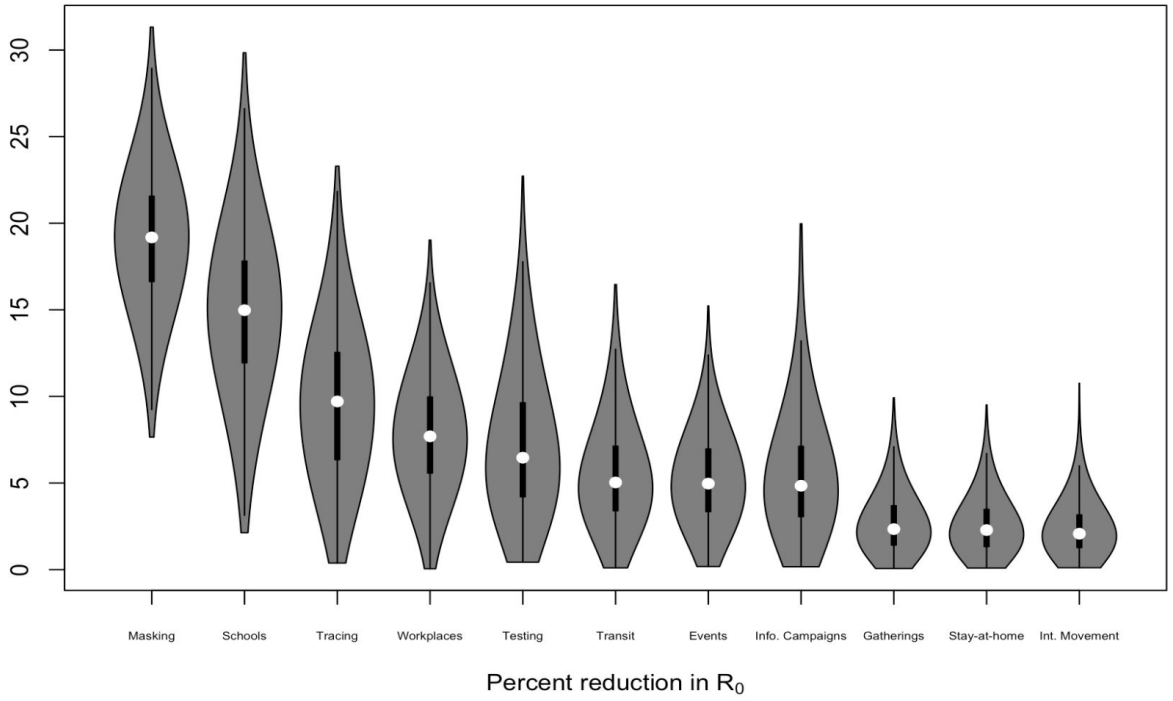


Figure 3.6: Posterior violin plots of global NPI effects.

Effects of individual NPIs

Figure 3.6 shows the pooled effect of each NPI, $\beta_u(k), k = 1, \ldots, p$, quantified as a percent reduction in R_0. Mask mandates are the most effective intervention, reducing R_0 by 19.2% (12.0%–26.2%). By comparison, Sharma et al. (2021) estimate that mask mandates reduced transmission rates by 12% (7%–17%) in the second wave in Europe. Based on data from 190 countries between January and April 2020, Bo et al. (2021) conclude that mask mandates were associated with a 15.1% (7.9%–21.8%) decline in transmission. Karaivanov et al. (2021) find a 22 percent weekly reduction in new COVID-19 cases due to mask mandates in Canada in the summer of 2020. Studying the 2020 spring wave in New York City, Yang, Shaff, and Shaman (2021) find that masking was associated with a 7% transmission reduction overall

and up to 20% reduction for people over age 65. Estimating the causal effects of a number of interventions in the US, Chernozhukov et al. (2021) demonstrate that masking policies were highly effective, leading to a reduction in the weekly growth rate of cases and deaths by more than 10 percentage points, with their conclusions holding robustly across model specifications; on the other hand, the effects of stay-at-home orders and business and school closures are much more uncertain. Qualitatively, our results are consistent with a number of other studies demonstrating the efficacy of mask mandates and the protective effects of face mask use (Greenhalgh et al., 2024; Jamison et al., 2021; Li et al., 2022; Lyu & Wehby, 2020; Rader et al., 2021; Talic et al., 2021).

Behind mask mandates, school closures are highly effective, reducing the transmission rate by 15.0% (5.6%–23.4%). By comparison, Brauner et al. (2021) and Banholzer et al. (2021) find that school closures led to 38% (16%–54%) and 17% (-2%–36%) transmission reductions in the first wave of 2020, respectively, and Sharma et al. (2021) estimate a 7% (4%–10%) transmission reduction due to school closures in the second wave. Studying influenza outbreaks, Cauchemez et al. (2008) found that school holidays led to a 20–29% transmission reduction among children with no detectable effect on transmission among adults. Qualitatively, our results are consistent with a number of other studies finding school closures to be one of the NPIs most effective in reducing transmission (Auger et al., 2020; Ferguson et al., 2006; Haug et al., 2020; Li et al., 2021; Liu et al., 2021; Markel et al., 2007; Stokes et al., 2022).

We are unable to rule out small effects for the remaining NPIs. Workplace closure reduced transmission by 7.7% (2.2%–14.0%). By comparison, Brauner et al. (2021), Sharma et al. (2021), and Banholzer et al. (2021) find that business closures led to a 27% (-3%–49%), 35% (29%–41%), and 18% (-4%–40%) reduction in R_0, respectively. The combination of social distancing measures[16] yields a 21% (13%–31%) reduction in R_0, which is on par with

[16]As in Section 3.2.4, we define social distancing measures as the combination of stay-at-home orders,

the individual effects of mask mandates, school closures, and the fear of infections reported above. Looking at individual social distancing measures, we estimate that stay-at-home orders reduced R_0 by a modest 2.3% (0.4%–6.4%), which is comparable to other estimates in the literature: Bodenstein et al. (2022) (who employ a more limited set of controls) report a 6.5% transmission reduction; Brauner et al. (2021) report a 13% (-5%–31%) reduction; and Banholzer et al. (2021) report a 4% (-6%–17%) reduction. In line with Banholzer et al. (2021), Brauner et al. (2021), Jamison et al. (2021), Li et al. (2021), Liu et al. (2021), and Stokes et al. (2022), we find that stay-at-home orders are among the least effective NPIs—which, in our case, include restrictions on gatherings and internal movement restrictions. There are a number of plausible explanations. For one, stay-at-home orders may yield smaller net reductions in R_0 due to increased household transmission when people spend more time at home. Additionally, the three weakest interventions—stay-at-home orders, restrictions on gatherings, and internal movement restrictions—were closer to voluntary rather than compulsory limitations on social mixing as implemented, since they were not strictly enforced. However, a number of other studies estimate large effects of strict gathering restrictions: Brauner et al. (2021) report a 42% (17%–60%) transmission reduction; Sharma et al. (2021) report a 26% (18%–32%) reduction; Banholzer et al. (2021) report a 37% (21%–50%) reduction; and Bo et al. (2021) report a 42.9% (41.6%–44.2%) reduction associated to social distancing measures more broadly.[17] Nevertheless, even with our relatively conservative estimates of the effects of social distancing measures, we find that they are cost-effective interventions in combination, as we demonstrate in Section 3.3.3.

Finally, our estimates of the effects of testing and tracing policies—which yield 6.5% (1.5%–16.2%) and 9.7% (2.2%–18.7%) reductions in R_0, respectively—allow for the possibil-

---

restrictions on gatherings, restrictions on internal movement, public information campaigns, public transit closures, and public event cancellations.

[17]We note that these are association studies based on observational data, i.e., they do not control for potential confounders.

ity of both small and large effects on transmission. Evidence on the effectiveness of contact tracing, in particular, is mixed. Rainisch et al. (2022) concluded that case investigation and contact tracing were effective in reducing transmission based on their estimates of the number of COVID-19 cases and hospitalizations averted by these measures in the US. Wang et al. (2022) find that testing and case isolation were effective, but the effect of contact tracing is marginal due to slow follow-up times in case investigation. They note that contact tracing can be more effective if follow-up is accelerated. In modeling studies, Hellewell et al. (2020) and Davis et al. (2021) find that contact tracing can be effective if carried out well—with the latter reporting a potential 15% reduction in R_0—but that its effectiveness is dependent on a number of epidemiological and implementation-related factors, including tracing coverage and speed. Nevertheless, even allowing for small effects, we find that testing and tracing are highly cost-effective NPIs owing to their low cost relative to other interventions.

3.3.3 Evaluating and optimizing costs

Baseline scenario

Figure 3.7 displays the results of our policy evaluation and optimization methodology detailed in Section 3.2.4 under the baseline cost scenario, which uses a cost function based on the medium value of the cost of workplace closure, the low VSCD (i.e., adjusted for the age pattern of COVID deaths), and the low cost of learning loss (i.e., assuming distance learning is at worst 90% as effective as in-person schooling) listed in Tables 3.2 and 3.3. The upper panels of Figure 3.7 display boxplots of the COVID death rate and per capita cost (including life costs) incurred by various policies across states. We consider the following five policy strategies: optimal control (OC), which is the strategy optimizing the cost function (3.5); full lockdown (Full), which assumes all NPIs are enforced at their strictest level for the entire year; observed (Obs.), the policy actually implemented; the observed policy minus school closure (Obs. - school); and the open policy (Open), which assumes no use of NPIs.

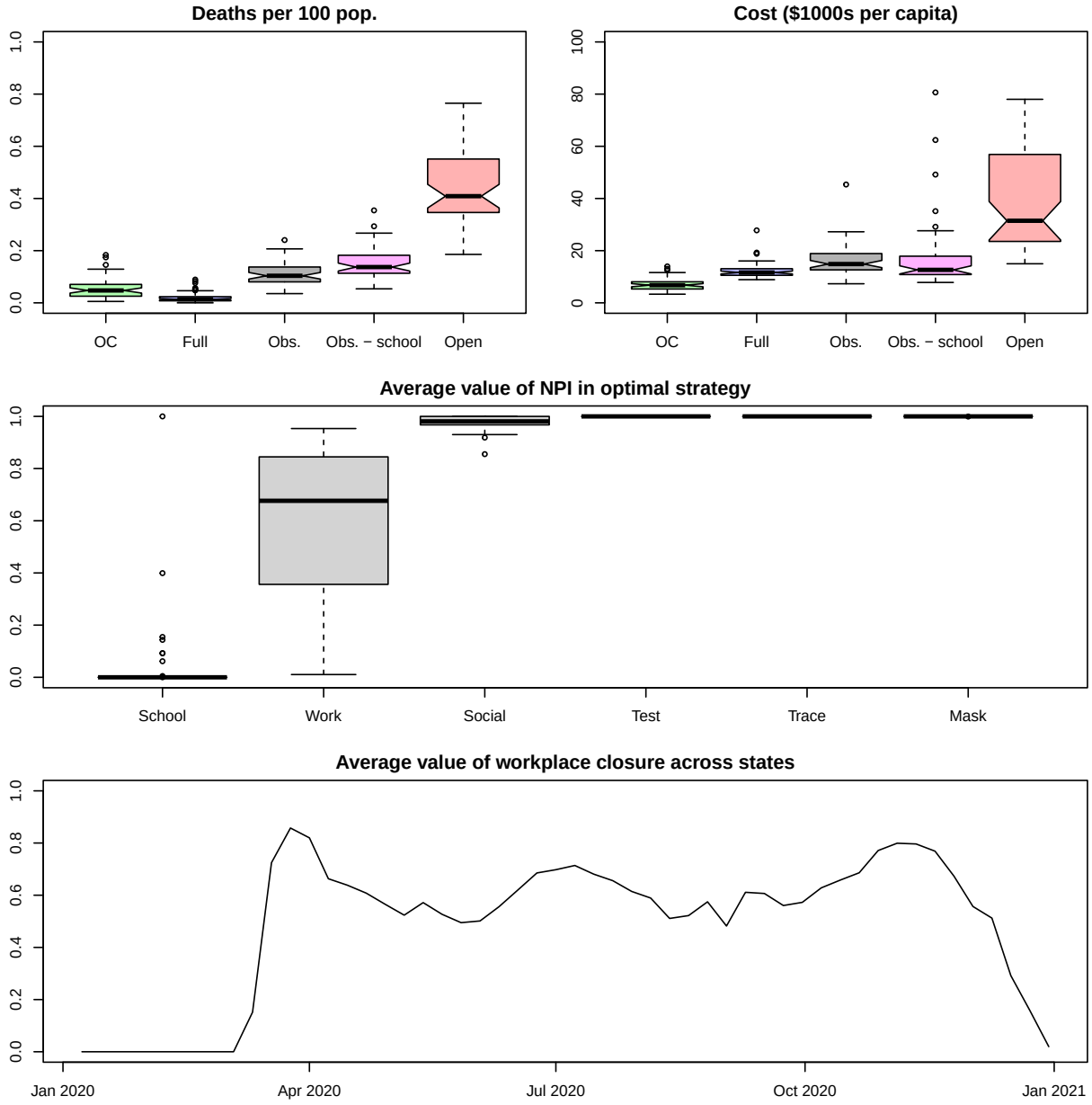


Figure 3.7: **Top and middle panels:** Boxplots of the following quantities across states: (i) posterior median of deaths per 100 population incurred by the optimal control (OC), full lockdown (Full), observed (Obs.), observed minus school closures (Obs. - school), and fully open (Open) policies; (ii) expected total cost in thousands of USD2020 per capita incurred by each policy; (iii) the average strength of each NPI in the optimal strategy. **Bottom panel:** the average strength of optimal workplace closures across states in each week.

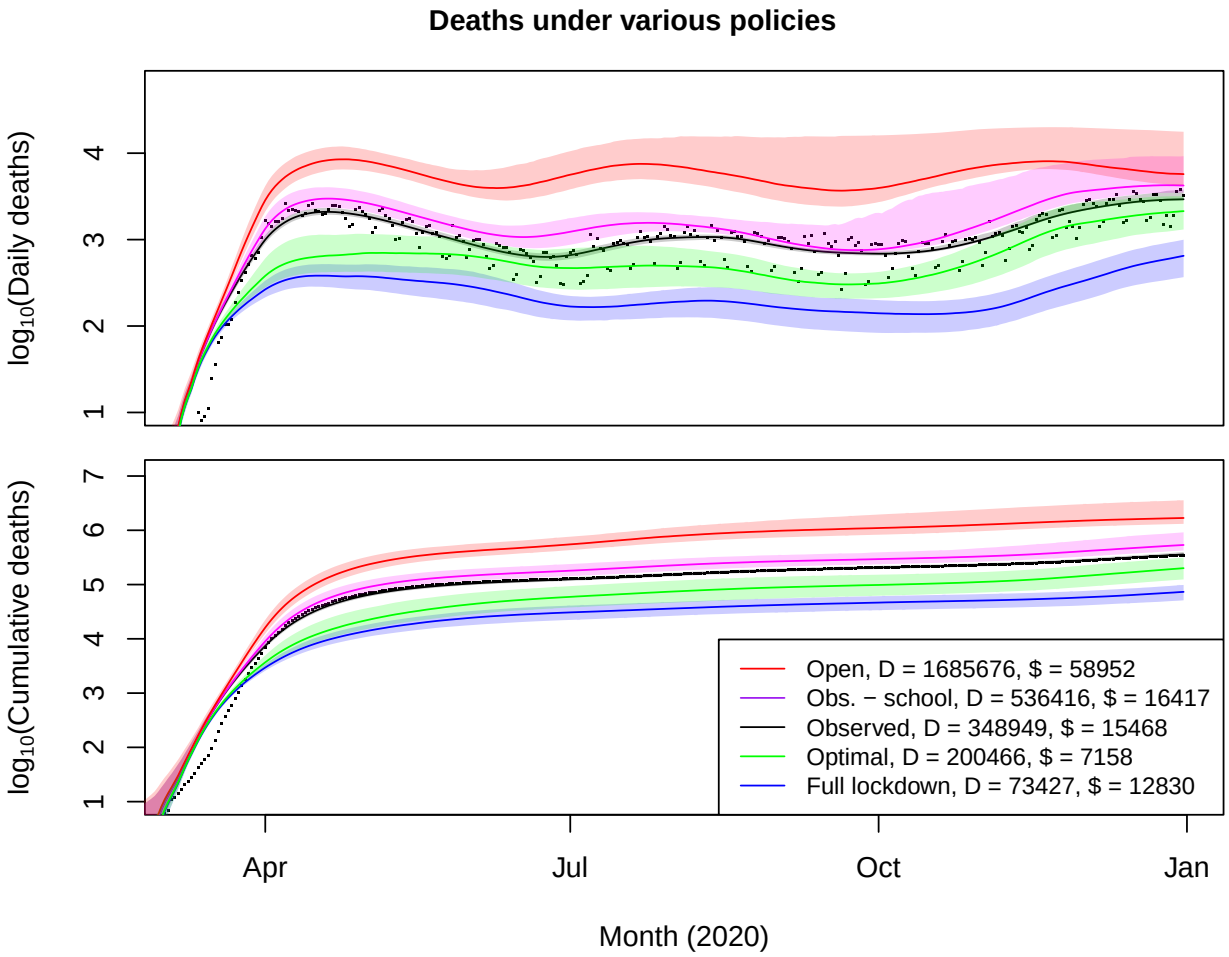


Figure 3.8: Posterior trajectories of daily deaths and cumulative deaths in the US under various strategies. The legend records the posterior median of the total deaths and expected total cost per capita in USD2020 incurred by each policy.

Full incurs the fewest deaths—as we would expect—followed by OC, then Obs., then Obs. - school, and finally Open. Regarding overall costs, OC is the least expensive policy (by definition), followed by Full, then Obs. and Obs. - school (which are approximately on par),

and finally Open. The middle panel of Figure 3.7 displays boxplots of the average strength of each NPI in the optimal control strategy over the year across states. An average strength of 1 implies that the intervention is implemented in its strictest sense for the entire year uninterrupted; an average strength of 0 implies that the intervention is never implemented at all.

Barring a few outliers, the OC strategies in each state are uniformly comprised of: consistent and strict use of social distancing measures, testing, tracing, and masking; no use of school closure; and moderate to strong use of workplace closure. The bottom panel of Figure 3.7 displays the average strength of workplace closure across states over time. Generally, the optimal strategy involves ramping up workplace closure to combat new waves of infections, with implementation peaking in response to the spring, summer, and fall waves of 2020. In Section 3.3.3 we explore the sensitivity of these results to plausible perturbations of the cost function. Our qualitative findings are robust across a range of scenarios with some qualifications.

Figure 3.8 displays the result of aggregating the state-level policy outcomes to the US as a whole. The open strategy incurs the most deaths and is the most expensive by far, with its cost arising entirely from infections. We estimate that 1.7 (1.4–4.5) million COVID deaths would have occurred in 2020 in the absence of public health interventions—similar to the 2.2 million projected by Ferguson et al. (2020). The burden of infections under the open strategy yields an expected cost per capita of $59,000 USD2020, which translates to a gross impact of $19.4 trillion—about 91% of US GDP in 2019 (U.S. Bureau of Economic Analysis, 2020). On the other hand, under full lockdown, we would have observed only 73,400 (47,700–103,400) COVID deaths and a cost to society of $12,800 per capita, which are surpassed by the 349,000 deaths and $15,500 per capita (or $5.1 trillion total) lost under the observed policy. Note, however, that full lockdown becomes more expensive than the observed policy and on par with the open strategy if we assume a high cost of learning loss. See Section

3.3.3. The posterior median mortality rate under full lockdown is $73,400/329.5 \approx 223$ deaths per million, which is about 60% of the COVID mortality rate observed in Canada in 2020 (Mathieu et al., 2020)

Although deaths under the observed policy minus school closure (Obs. - school) substantially exceed those under Obs., the expected costs to society of each policy are comparable. Hence, conditional on the other NPIs implemented, the decision to close schools presents a marginal trade-off between the learning of students and the health of those most vulnerable to COVID. The extended school closures that were enacted across the country in 2020 prioritized the latter in lieu of the former. While our model estimates that they saved approximately 190,000 lives, this came at the expense of $2 trillion in lost learning. However, this trade-off was not inevitable. Under the optimal policy, which involves no closure of schools in 2020 beyond the usual 16 weeks of break, we incur 200,500 (132,800–497,000) deaths at a cost to society of $7,200, with most of the savings relative to the observed policy stemming from the cost of infections and school closure. We find that more timely, stringent, and enduring use of other measures—social distancing, testing, tracing, and masking, along with reactive workplace closures—would have been sufficient to limit COVID mortality at or below what was observed without incurring profound learning loss.

We can compare our estimates of the gross impacts of COVID-19 and NPIs to others in the literature. Bruns and Teran (2022) and Cutler and Summers (2020) project the total cost of the pandemic in the US over its full duration at $16 trillion, which is slightly more than three times our estimate of $5.1 trillion in losses observed during 2020. Flaxman et al. (2020) estimated that NPIs averted 3.1 (2.8–3.5) million deaths up to May 2020 in 11 countries totaling 375 million population. Ferguson et al. (2020) predicted that NPIs would prevent 1.1 million deaths in the US over the course of the pandemic. Greenstone and Nigam (2020) projected that moderate social distancing would save 1.7 million lives in the US by October 1, with mortality benefits of $8 trillion, or $24,000 per capita, based

on the same VSCD used in our baseline scenario. They note that the vast majority of the monetized benefits of social distancing accrue to people age 50 or older. More conservatively, Eichenbaum et al. (2021) find that containment policies, if implemented optimally, would save about half a million lives in the US based on low values of key epidemiological parameters—specifically, they use an IFR of 0.5% and an R_0 of 1.45 for their baseline model. Thunström et al. (2020) estimated that social distancing in the US would save 1.24 million lives at a cost of $7.2 trillion in lost GDP, which implies that social distancing measures would yield net losses for any VSCD below $7.2/1.24 \approx 5.8$ million dollars. To the contrary, we find that NPIs (and social distancing measures in particular) are cost-effective for a VSCD of $4.5 million. Assuming a vaccine arrives (stochastically) after a year to end the pandemic, Farboodi et al. (2021) estimate the per capita cost of the optimal policy at $8,100, similar to our $7,200. However, they find that the *laissez-faire* equilibrium (i.e., in the absence of government intervention) would only incur a cost of $12,700 per person. Undertaking a cost-benefit analysis of confinement policies targeted toward mitigation and (strict) suppression, Gollier (2020) finds that both strategies incur a total cost—combining economic and life costs—equating to 15% of annual GDP, or about $10,000 per capita, which is comparable to our estimates of the total costs of the various containment strategies in Figure 3.8. Jones et al. (2021) estimate that, under an optimal social distancing policy, GDP declines by 12% and 0.17% of the population (about 560,000) die from COVID in the first 26 weeks of the pandemic. In a simulation study, Keogh-Brown et al. (2020) estimate that containment strategies in the UK to suppress COVID-19 through the end of 2020 would incur health costs of 1.7% of GDP and economic costs of 29.2% of GDP, with 7.3 percentage points coming from workplace absenteeism of parents affected by school closure and 21.9 percentage points from business closure. Their total cost is comparable to our estimate for the U.S. in 2020 (i.e., under the observed policy), which is 23% of GDP. However, relative to the results of Keogh-Brown et al. (2020), we find that health impacts are a much larger portion of the

total, costs related to business closure are much smaller, and costs related to school closure are somewhat larger.

Incremental cost-effectiveness ratios (ICERs) in infectious disease

The incremental cost-effectiveness ratio (ICER) is a quantity widely used in the economic evaluation of health interventions. The ICER is defined as the monetary cost of an intervention divided by the benefit it produces (as measured by some target outcome) relative to a baseline, with a larger ICER often (mis)interpreted to mean that an intervention is less cost-effective (Paulden, 2020). In public health and infectious disease, outcomes of interest include the number of quality-adjusted life years (QALYs), disability-adjusted life-years (DALYs), infections, hospitalizations, or deaths averted by the intervention. Despite their ubiquity in the health economics literature, ICERs can be difficult to interpret and are often defined, calculated, and reported inconsistently, which severely limits their practical value in cost-effectiveness analysis (Paulden, 2020; Weinstein, 1990). As a result, ICER estimates for the same strategy can vary by orders of magnitude across studies and rankings of interventions based on their reported ICERs can yield counter-intuitive results, as evidenced by a recent systematic review of economic evaluations of COVID-19 interventions (Podolsky et al., 2022). For example, Podolsky et al. (2022) find that the median ICER of school closure across studies in their review is exceeded by that of vaccination, testing, facial covering, and stay-at-home policies.[18] Furthermore, the median ICER of mask mandates exceeds that of school closure by an order of magnitude, and the median ICER of stay-at-home orders exceeds that of school closure by nearly 3 orders of magnitude. To address these issues, some have argued instead for the use of a policy's net benefit in decision-making (Craig & Black, 2001; Paulden, 2020; Stinnett & Mullahy, 1998). We take a similar approach in Section 3.3.3,

[18]This may be due to the fact that many cost-effectiveness analyses considering school closure fail to account for costs associated to student learning loss, as we discuss in Section 3.1.

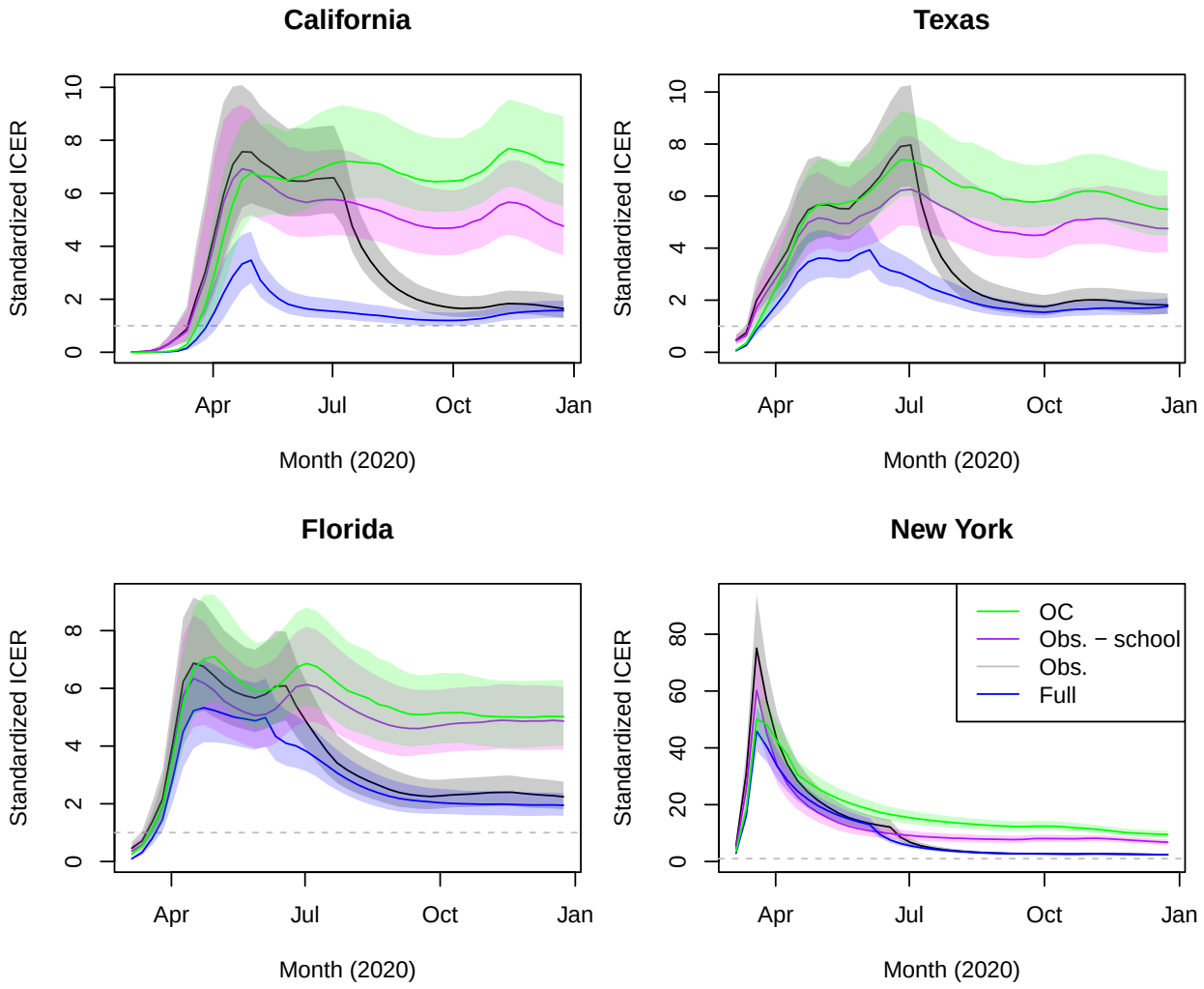


Figure 3.9: Posterior median and 50% credible intervals for the cumulative standardized ICER of various policies in the four most populous states.

reporting the expected total cost of various strategies, which facilitates straightforward comparison of their cost-effectiveness. Nevertheless, our methodology can shed light on the appropriate use of ICERs in the economic evaluation of interventions targeting infectious disease transmission.

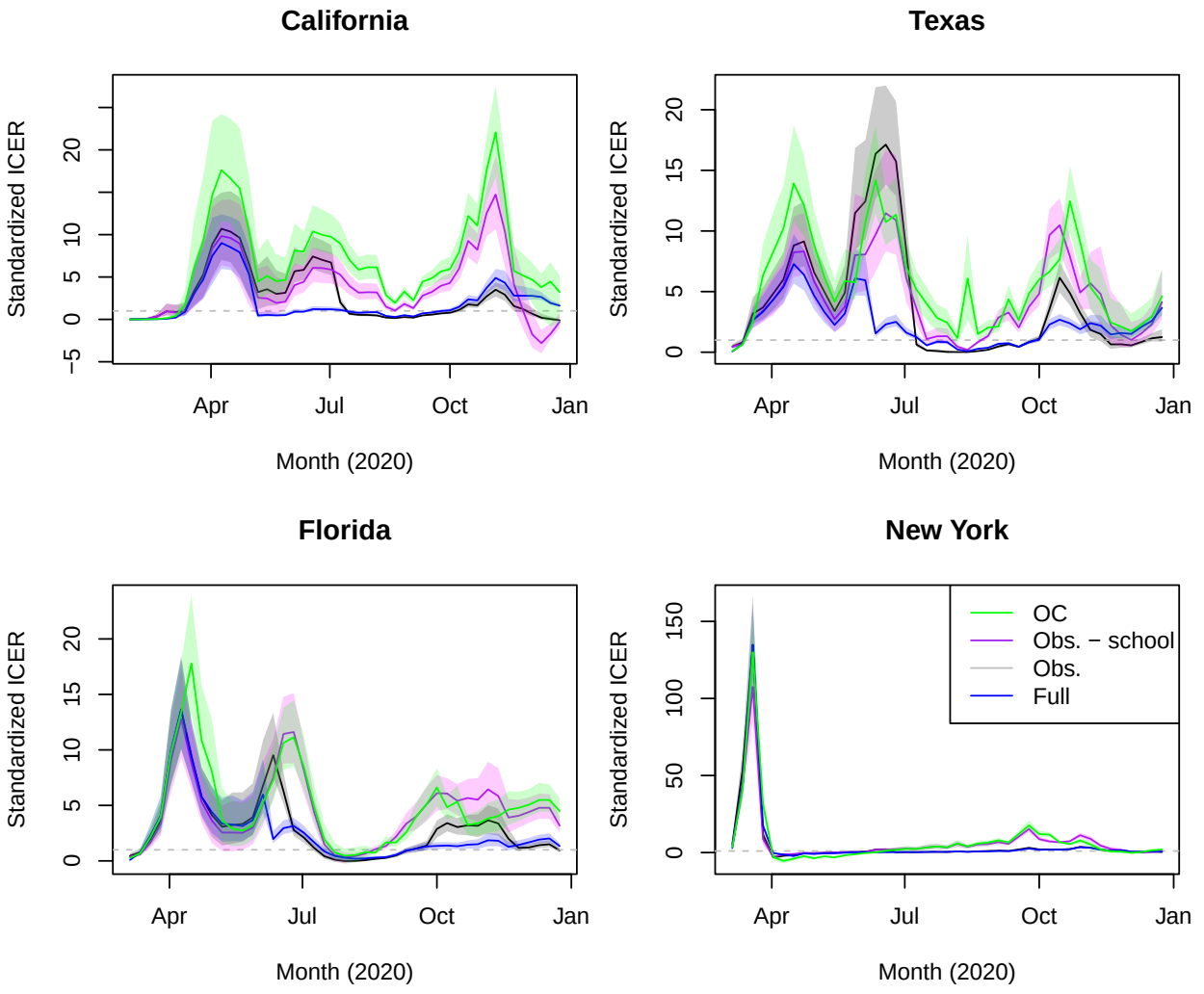


Figure 3.10: Posterior median and 50% credible intervals for the weekly standardized ICER of various policies in the four most populous states.

We argue that ICERs in the context of infectious disease interventions should be: data-driven and account for uncertainty; not based solely on (poorly calibrated, overly simplistic, or deterministic) simulation models; reported based on well-defined interventions[19]; and

[19]Relevant details include the duration and strictness of implementation and the strength of adherence

calculated using the intervention's (causal) effect on the effective reproduction number $R_e(t)$, the relevant quantity governing infectious disease transmission, which is interpreted as the expected number of secondary cases resulting from an infection and can be estimated from clinical data in real time. Related to this last point, we note that, for a given intervention, the ICER is a time- and context-specific quantity.[20] Therefore, in calculating and reporting the ICER, our results should account for or be as invariant as possible to contextual factors. We echo Prager et al. (2017), who highlight the "'...importance of including a broader set of causal factors to achieve more accurate estimates of the total economic impacts of not just pandemic influenza but biothreats in general."

If we consider a blanket intervention that affects the population at large (e.g., social distancing measures), the intervention's reduction in $R_e(t)$ multiplied by the estimated size of the infectious population gives the number of infections averted by the intervention on day t, which forms the denominator of the ICER. On the other hand, if we consider a targeted intervention (e.g., case isolation), the number of infections averted can be estimated as the reduction in $R_e(t)$ multiplied by the number of treated subjects. Finally, the numerator of the ICER is the cost of implementing the intervention on day t.

To demonstrate, Figures 3.9 and 3.10 exhibit cumulative and weekly standardized ICERs of various strategies relative to the open policy (Open)—in which no NPIs are used—over time in the four most populous US states. We calculate these standardized ICERs as follows. For an NPI policy u implemented on a specific day t, let

$$R_e(u,t) := R_0(u,t)S(t)$$

denote the effective reproduction number on day t under policy u, where (in an abuse of

---

to the intervention.

[20]This is an important concern when working with ICERs. Outcomes, such as infections averted by an intervention, depend also on the implementation of other interventions (their timing, strength, duration) as well as other factors affecting disease transmission (e.g., the duration of the pandemic, the baseline R_0 value, other epidemiological parameters, voluntary social distancing and protective measures, exogenous shocks, new variants, etc.).

notation), $R_0(u,t)$ is the basic reproduction number defined by our NPI regression model (3.3) and $S(t)$ is the susceptible fraction of the population. The number of infections averted by the policy u on day t relative to the open strategy is then

$$\nu_a(t) := NI(t)(R_e(\mathbf{0},t) - R_e(u,t)),$$

where N is the population size, $I(t)$ is the infectious fraction of the population, and $R_e(\mathbf{0},t)$ denotes the effective reproduction number under no interventions. In a given period of days $[T_1,T_2]$, the ICER of a policy $\mathbf{u} = \{u(t) : t \in [T_1,T_2]\}$ is then

$$\text{ICER}_{\mathbf{u}}(T_1,T_2) = \frac{c_{\text{NPI}}(\mathbf{u}) \cdot N}{\sum_{t=T_1}^{T_2} \nu_a(t)},$$

where $c_{\text{NPI}}(\mathbf{u})$ is the per capita cost of the strategy, as defined in Section 3.2.4. Finally, to define the standardized ICER plotted in Figures 3.9 and 3.10, we convert infections to their monetary value using the cost of an infection c_ν—defined in Section 3.2.4—and take the reciprocal, such that the standardized ICER reports the ratio of the value of infections prevented to the cost of the intervention, which is a dimensionless quantity:

$$\text{SICER}_{\mathbf{u}}(T_1,T_2) := \frac{\sum_{t=T_1}^{T_2} c_\nu \nu_a(t)}{c_{\text{NPI}}(\mathbf{u}) \cdot N}. \tag{3.6}$$

Based on (3.6), in terms of net costs, a policy $\mathbf{u}$ relative to no intervention in time period $[T_1,T_2]$: breaks even if SICER $= 1$; produces a net benefit if SICER > 1; and produces a net loss if SICER < 1.

We now return to Figure 3.9, which displays the cumulative standardized ICER over time, $\{\text{SICER}_{\mathbf{u}}(1,t)\}_{t=1}^{T}$ for various policies $\mathbf{u}$. We see that, relative to Open, the different containment strategies are comparable in terms of cumulative cost-effectiveness early in the pandemic, with Obs. and Full becoming less cost-effective (but still producing a net benefit by the end of 2020), mainly due to the cost of school closures exceeding 16 weeks. The optimal control strategy OC is the most cost-effective by the end of the year, with the

SICER ranging from about 5–10 across states, while Obs. - school is a close second. Note that, while OC and Obs. - school are nearly equally cost-effective by the end of 2020 in Florida, this does not imply that Obs. - school is also an optimal policy—this can only be determined by looking at the net benefit of each policy. Figure 3.10 displays the standardized ICER of each policy in each week (conditional on that policy also being implemented in all weeks prior) relative to Open. Results are similar: OC and Obs. - school are similarly cost-effective and consistently more cost-effective than Obs. and Full; all containment strategies satisfy SICER > 1 for most of the year, implying that they produce net benefits relative to no intervention.

The optimal control strategy is determined by minimizing aggregate costs accrued over the year. We can solve this problem *ex post*—after we have observed the pandemic play out—but, in principle, we cannot derive an optimal policy *ex ante*, since we cannot see the future. Indeed, while SIR models have demonstrated remarkable utility in helping us understand infectious disease (Kermack & McKendrick, 1927), modeling studies carried out early in the pandemic (e.g., Ferguson et al. (2020)) generally failed to predict the complex and stochastic dynamics of SARS-CoV-2 (e.g., multiple waves, super-spreader events) depicted in Figure 3.4. As we note in Section 3.3.2, even when we incorporate the effects of interventions into the model, we fail to explain a substantial portion of the temporal variation in transmission rates. At its face, this may seem like a disheartening realization. However, the trends in Figures 3.9 and 3.10 have important and encouraging implications for decision-making during pandemics. Specifically, they show that the OC policy—a relatively simple combination of testing, tracing, masking, social distancing, and reactive workplace closure—was consistently highly cost-effective on a weekly basis throughout the first year of the pandemic. This implies that, had we reasonable ballpark estimates of the costs and effects of interventions to work with early on, we could have determined a nearly-optimal strategy by choosing, at each point in time, the policy that greedily minimized the cost incurred in the next time step.

In our context, we could not predict long-run transmission rates and therefore we did not need to; the myopic strategy would have produced a nearly-globally-optimal solution. A similar phenomenon has been documented in the control theory community (Recht, 2024): the performance of control algorithms can be highly sensitive to modeling errors. Hence, if a model is misspecified, or if we can only poorly understand the behavior and evolution of a control system, simple algorithms tend to be more robust.

We turn now to the cost-effectiveness of school closure in particular. In Section 3.3.3, we estimate that, relative to Obs. - school, the observed policy saved about 190,000 lives with the cost of school closures amounting to $2 trillion, yielding an ICER of about $10.5 million per death prevented. Our estimate is similar to that reported in the systematic review of Juneau et al. (2022), who find that school closures during the 2009 H1N1 influenza pandemic cost $9.86 million per death prevented, which implies that the cost of preventing a death (due to H1N1 or COVID-19) is on par with the VSL—taken to be $10.63 million, in line with Robinson et al. (2021), in our study. As such, in our sensitivity analysis in Section 3.3.3, we find that optimal policies involve school closure to some degree when the VSCD is assumed equal to the VSL and the cost of learning loss remains at the low value used in our baseline scenario.

Qualitatively, our results are consistent with the findings of Juneau et al. (2022) in other ways as well. Specifically, they find that: testing, tracing, and masking are among the most cost-effective measures; workplace and school closures are effective but costly, and hence the least cost-effective interventions; combinations of NPIs are more cost-effective than single interventions; and NPIs are more cost-effective when implemented early. While Juneau et al. (2022) conclude that school closure is among the least cost-effective interventions based on the ICER, the value of this finding is limited for a number of reasons.

Firstly, as Paulden (2020) notes, it is difficult to decide based on the ICER alone if an intervention is cost-effective in a given setting, i.e., if it would be implemented in an optimal

policy in conjunction with other interventions and other external factors. Indeed, as noted above, the cost-effectiveness of an NPI varies over time and across contexts, particularly as prevalence varies.[21] As such, calculating and ranking the ICERs of various interventions is not sufficient to determine which strategies are cost-effective (Paulden, 2020). However, as we demonstrate in Section 3.3.3, we can establish that school closure is not cost-effective by deriving the optimal policy, which frames the analysis in terms of the expected total cost—or, equivalently, the net benefit—of a policy. Our methodology goes beyond quantifying the cost-effectiveness ratios of different policies by determining which NPIs should have been used and when.

Secondly, the results of Juneau et al. (2022) are based on analysis of school closures during the 2009 H1N1 pandemic. The biology and epidemiology of the H1N1 flu differed from SARS-CoV-2 in important ways. In particular, it was evident early in the COVID-19 pandemic that SARS-CoV-2 was more virulent and more transmissible than H1N1, which left open the possibility that school closure would be cost-effective in combating SARS-CoV-2 transmission in 2020 despite its apparent lack of cost-effectiveness in 2009, as noted by Pasquini-Descomps et al. (2017) and Xue et al. (2012). Similarly, studying the cost-effectiveness of interventions in response to outbreaks of influenza, gastroenteritis, and chickenpox in France, Adda (2016) finds that school closures are not cost-effective, but they "...would become beneficial for epidemics characterized by a slightly more deadly strain." Indeed, Dauelsberg et al. (2024), Kelso et al. (2013), Milne et al. (2013), Perlroth et al. (2010), and Xue et al. (2012) find that extended school closures are cost-effective for severe pandemics. We note that Dauelsberg et al. (2024), Kelso et al. (2013), Milne et al. (2013), and Perlroth et al. (2010) do not account for costs associated to student learning loss and Xue et al. (2012) do not consider interventions other than school closure as available tools in their model. As we show in Section 3.3.3, in the context of COVID-19, school closures are not cost-effective.

[21] This is one factor explaining why earlier is better with regard to the timing of NPIs: as population immunity grows, NPI implementation yields diminishing returns.

Sensitivity analysis of optimal control strategies

Figures 3.11 and 3.12 display the results of our sensitivity analysis. We vary the value of a statistical COVID death (VSCD), the cost of learning loss, and the cost of workplace closures across plausible ranges, which are given in Tables 3.2 and 3.3. As noted above in Section 3.2.4, we do not vary the cost of social distancing measures as we are primarily interested in assessing the robustness of the optimal strategy and the relative costs of various policies rather than variation in the total cost incurred by each policy. We also omit sensitivity analysis for the costs of testing, tracing, and masking as they are highly cost-effective interventions under any reasonable variation in their costs. Indeed, these measures are all at least somewhat effective in reducing transmission and extremely cheap compared to infections, school and workplace closures, and social distancing mandates.

Figure 3.11 exhibits boxplots of the average value of each NPI in the optimal strategy across states. We see that school closures are only implemented assuming both a high VSCD equal to the VSL ($10.63 million USD2020)—which does not adjust the VSCD for the age profile of COVID mortality—and a low cost of learning loss (9% of GDP per 0.33 years)—which assumes that distance learning is 90% as effective as in-person schooling—which are shown in the upper left panels. Moreover, when the cost of workplace closure is set to our low scenario (2% employment rate decrease), school closures are still implemented only sparingly, with a median across states of 0.15 school-years of closure. Across scenarios, workplace closures are implemented fairly consistently, but their optimal duration decreases with increasing cost. These conclusions are comparable to the results of Barrot et al. (2024), who find that the cost-effectiveness of business closures (i.e., whether they produce a net benefit or loss) is sensitive to modeling assumptions and particularly the assumed value of a life-year.

Figure 3.12 shows boxplots of the costs of various policies across states. Notably, when the cost of learning loss is high (right column), the total cost of the observed policy (i.e.,

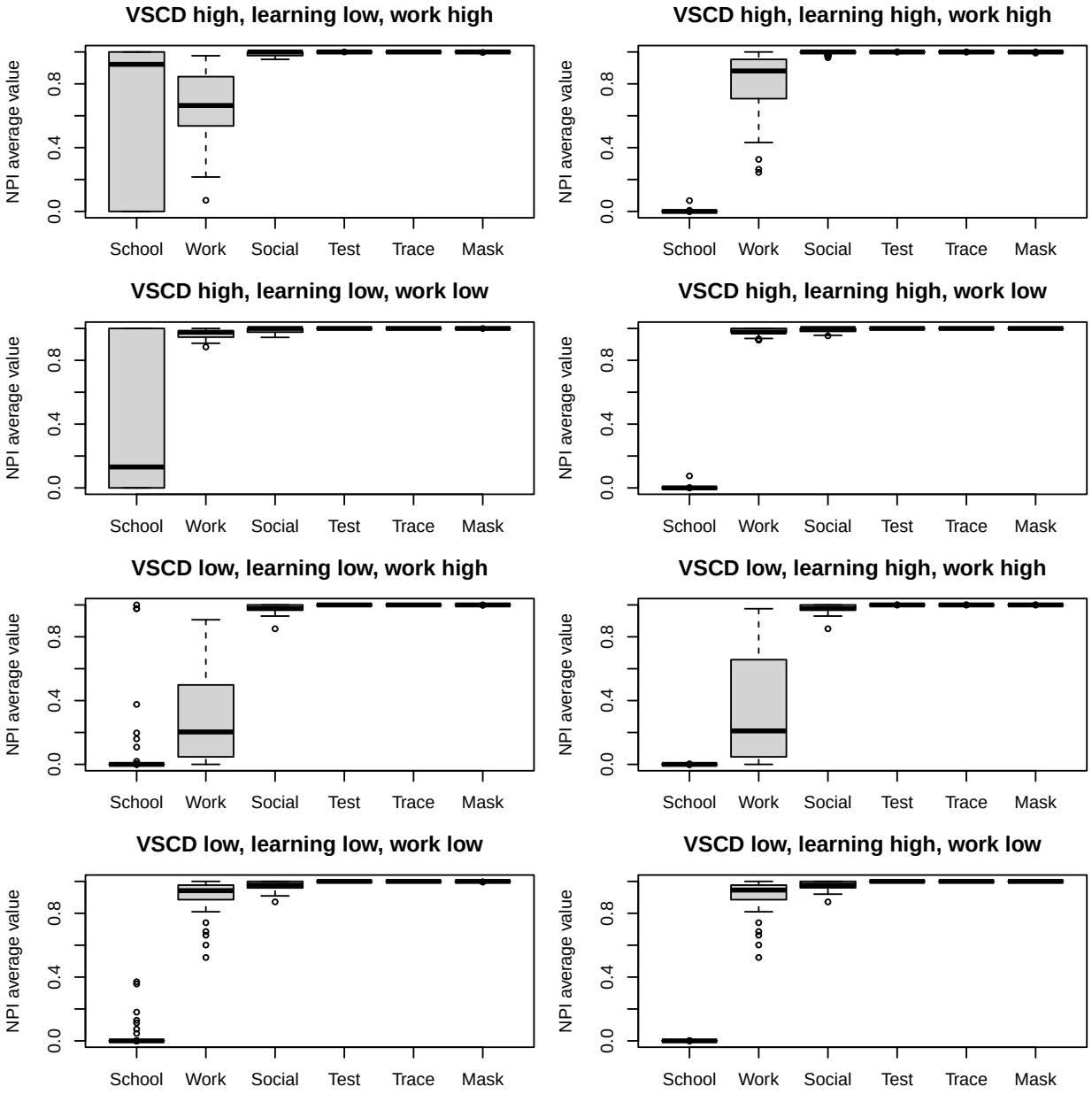


Figure 3.11: Sensitivity analysis for the optimal control results. Boxplots of the average value of the optimal NPI policy over the year in each state.

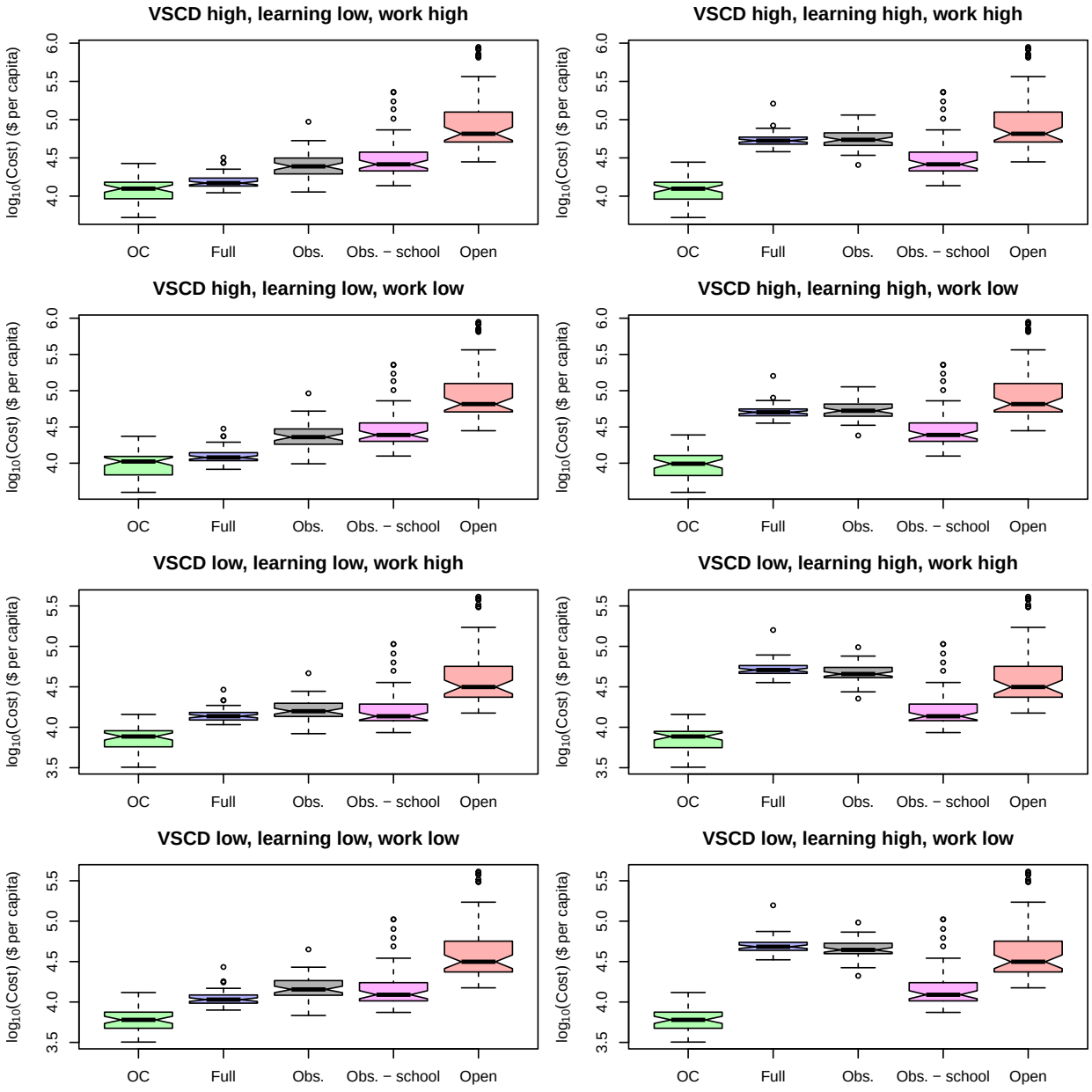


Figure 3.12: Sensitivity analysis for the costs of various policies. Boxplots of the log-scale total cost in USD2020 per capita incurred by the optimal control (OC), full lockdown (Full), observed (Obs.), observed minus school closures (Obs. - school), and fully open (Open) policies across states.

the one that was actually implemented) minus school closures (denoted Obs. - school) is substantially lower than the observed (Obs.) and full lockdown (Full) policies in almost every state. If we further assume that the VSCD is low (bottom right panels), Full exceeds the cost of Obs., which exceeds the cost of the fully open policy (Open) in many states. If both VSCD and the cost of learning loss are low (bottom left panels), Obs. - school becomes cheaper than Obs. and on par with Full in many states. In most settings, Full is cheaper than Obs., which is cheaper than Open.

3.4 Discussion

We have developed a statistical decision framework in order to conduct a cost-effectiveness analysis of non-pharmaceutical interventions in the U.S. during COVID-19. While the use of NPIs poses health, economic, and social trade-offs, it is not a zero-sum game. As others have noted (Demirgüç-Kunt et al., 2021; Fernández-Villaverde & Jones, 2020; Kaplan et al., 2022), appropriately implemented restrictions can simultaneously limit deaths and the aggregate costs to society incurred during a pandemic. Our methodology enables us to derive optimal NPI strategies, which consist of timely, enduring, and stringent use of testing, tracing, and masking policies, social distancing measures, and reactive workplace closure, with no closure of schools.

This last finding is salient as schools were closed for extended durations in the U.S. and in many other countries throughout the pandemic. Growing evidence suggests that the impacts on school children are substantial and long-term (Mervosh et al., 2024; UNICEF et al., 2021). As we show in Section 3.3.3, our conclusion that school closure is not cost-effective is robust to plausible variation in the cost function (which, as we note in Section 3.2.4, conservatively accounts for the cost of school closure). Furthermore, given that we estimate school closure to be one of the interventions most effective in reducing transmission, our results would not be easily overturned based on different modeling assumptions. If we have over-estimated the

effect of school closure on transmission reduction for most of 2020, getting closer to the truth would only strengthen our findings. Relative to the first COVID wave in Europe (Brauner et al., 2021), Sharma et al. (2021) find that school closure was substantially less effective in reducing transmission, with their point estimate of the effect (7%) amounting to about half of ours (15%). Sharma et al. (2021) speculate that the effect attenuated from the first to the second wave because many schools in Europe reopened without substantial increases in transmission. Some have argued that, with adequate health protocols in place (Lordan et al., 2020; Rice et al., 2020; Viner et al., 2021), U.S. schools that remained closed through the 2020–2021 academic year could have resumed in-person learning safely (Mervosh et al., 2024).

We note that, for practical purposes and due to lack of available data, our model of SARS-CoV-2 transmission does not account for a number of complexities. We do not explicitly account for the age structure of a state's population and its infections, although these are reflected in the infection fatality rate (IFR) estimates used in our model based on Irons and Raftery (2021). As such, our reported effects of NPIs on transmission and costs associated to infections and NPIs should be interpreted as aggregate measures.

We do not model state-level hospital capacity and potential excess costs or deaths arising from an overwhelmed medical system. In principle, doing so would serve to increase the cost associated to infections. However, we note that estimates of the IFR in England, which experienced COVID death rates similar to the U.S. in 2020 (Mathieu et al., 2020), are fairly constant over the year (Eales et al., 2023), which suggests that COVID mortality outcomes—the dominant term in the cost of infection—were not highly sensitive to fluctuations in the burden on hospitals.

As noted in Section 3.1, we do not account for mental health costs arising from lockdowns and from the fear of infection, which may be substantial (Bruns & Teran, 2022; Cutler & Summers, 2020). While there is some recent work estimating the causal effects of stay-at-

home orders and school closure on mental health outcomes (Ferwana & Varshney, 2024), to the best of our knowledge, the effects of other relevant exposures, including workplace closure, other NPIs, and the pandemic itself, have not been ascertained. Relatedly, we are not considering environmental or health-related costs or benefits concerning pollution, emissions, and traffic injuries resulting from reduced industrial activity, vehicle travel, and energy consumption during the pandemic due to mitigation policies or other factors.

Given the highly correlated implementation of NPIs and that we are already accounting for spatial heterogeneity of their effects in our hierarchical model, we do not also model temporal variation in NPI effects (e.g., arising from "pandemic fatigue"), which has been documented in some studies (Ge et al., 2022; Petherick et al., 2021; Sharma et al., 2021), as it would be difficult to identify from the data. As such, our estimates of the effects of NPIs on viral transmission should be interpreted as an average effect over the year. While Ge et al. (2022) find that the overall effect of NPIs in Europe increased over time in 2020, Petherick et al. (2021) observe increasing use of masks but declining adherence to physical distancing measures across countries over the year. As noted above, Sharma et al. (2021) estimate a reduced effect of school closure in the second wave in Europe, which comports with the findings of Petherick et al. (2021). In relation to our results, increasing efficacy of masking policies and decreasing efficacy of school closure would not overturn the conclusions that mask mandates are highly cost-effective whereas school closures are not. However, substantial decreases in the efficacy of workplace closure and social distancing measures could affect their appearance in optimal mitigation strategies.

Finally, as we model viral spread in the U.S. states independently of each other, we do not account for spillover effects of intervention policies between states, which may play a role in overall trends in transmission (Holtz et al., 2020).

Chapter 4

CAUSALLY SOUND PRIORS FOR BINARY EXPERIMENTS

We introduce the BREASE framework for the Bayesian analysis of randomized controlled trials with a binary treatment and a binary outcome. Approaching the problem from a causal inference perspective, we propose parameterizing the likelihood in terms of the **b**aseline **r**isk, **e**fficacy, and **a**dverse **s**ide **e**ffects of the treatment, along with a flexible, yet intuitive and tractable jointly independent beta prior distribution on these parameters, which we show to be a generalization of the Dirichlet prior for the joint distribution of potential outcomes. Our approach has a number of desirable characteristics when compared to current mainstream alternatives: (i) it naturally induces prior dependence between expected outcomes in the treatment and control groups; (ii) as the baseline risk, efficacy and risk of adverse side effects are quantities commonly present in the clinicians' vocabulary, the hyperparameters of the prior are directly interpretable, thus facilitating the elicitation of prior knowledge and sensitivity analysis; and (iii) we provide analytical formulae for the marginal likelihood, Bayes factor, and other posterior quantities, as well as exact posterior sampling via simulation, in cases where traditional MCMC fails. Empirical examples demonstrate the utility of our methods for estimation, hypothesis testing, and sensitivity analysis of treatment effects.

This chapter is based on the preprint "Causally sound priors for binary experiments" (Irons & Cinelli, 2023).

4.1 Introduction

Randomized controlled trials (RCTs) form the cornerstone of scientific research across numerous disciplines. In their most basic form, these trials compare the occurrence of an

adverse (or favorable) outcome between treatment and control groups. This is particularly evident in a drug or vaccine trial, in which the efficacy of an intervention is established by comparing the number of individuals who die or develop a disease in each arm of the study. We refer to this type of study design as a "binary experiment," wherein each participant is subjected to either a treatment or a control condition (a binary exposure), and we observe either the presence or absence of the adverse effect of interest (a binary outcome).

If participants of the trial are independent draws from a common (super-)population, statistical inference in binary experiments amounts to what is perhaps the simplest of tasks in statistics—the comparison of two binomial proportions. Indeed, from a Bayesian perspective, inference on the parameter of a binomial distribution dates back to at least as early as the origins of Bayesian inference itself, as evidenced by the seminal works of Bayes (1763) and Laplace (1774). The task comprises specifying a joint prior distribution for both binomial parameters, and computing the posterior distribution (or Bayes factors) of (relevant contrasts of) such parameters (e.g., the risk difference, or the risk ratio). Yet, perhaps surprisingly, despite this long tradition, their widespread occurrence in the sciences, and the apparent simplicity of the inferential task, mainstream approaches for prior specification in the analysis of binary experiments have several shortcomings.

As reviewed in Agresti and Min (2005) and Dablander et al. (2022) (and also evident from perusing popular textbooks[1]) the two predominant approaches for the Bayesian analysis of binary experiments consist of: (i) assigning independent beta priors to each of the binomial proportions, which are conjugate priors to the (also independent) binomials comprising the likelihood; and, (ii) what is essentially a logistic regression, i.e., applying a logit transformation to the binomial proportions, and assigning Gaussian priors to the average log odds and the log odds ratio. For all their popularity, these two approaches are unsatisfactory in several ways. For example, in the first case, the assumption of prior independence of the two

[1]See, e.g., Gelman et al. (1995), Kruschke (2014), and McElreath (2020).

proportions is often not credible—e.g., in most settings, one expects that learning about the mortality rate in the control group should inform our beliefs about the mortality rate in the treatment group. Moreover, while the logit approach addresses the problem of prior dependence, it does so at the sacrifice of clarity and interpretation—odds ratios are notoriously difficult to understand (Davies et al., 1998), thus hindering the utility of this approach for prior elicitation and sensitivity analysis.

In this paper we demonstrate how causal logic can be used to address these challenges. Approaching the problem from a causal inference perspective, we first propose parameterizing the likelihood in terms of three clinically meaningful counterfactual quantities: the baseline risk, efficacy, and risk of adverse side effects (BREASE) of the intervention. We then propose a flexible, yet intuitive and tractable jointly independent beta prior distribution on these parameters, which we show to be a generalization of the Dirichlet prior on the joint distribution of potential outcomes. Our approach has a number of desirable characteristics: (i) it naturally induces prior dependence between the two binomial proportions of the treatment and control arms of the study; (ii) as the baseline risk, efficacy and risk of adverse side effects are quantities familiar to clinicians, the hyperparameters of the prior are directly interpretable, thus facilitating the elicitation of prior knowledge and sensitivity analysis; and (iii) we derive analytical formulae for the marginal likelihood, Bayes factor, and other posterior quantities, as well as exact posterior sampling via simulation, in cases where traditional MCMC fails.

Related literature. When framed in the language of potential outcomes, causal inference can be seen as a missing data problem. Thus, our analysis is most closely related to the literature on contingency tables with missing or incomplete observations on certain cell counts. In fact, our proposed prior can be shown to induce a *generalized* Dirichlet distribution on the joint distribution of potential outcomes. This distribution has been studied in the 1970s and 1980s (Antelman, 1972; Dickey et al., 1987; Dickey, 1983; Kaufman & King, 1973), though

mostly in the context of survey sampling.[2] Perhaps due to the intractability of the integrals, the difficulty in interpretation of the original generalized Dirichlet parameterization, and the missing connection to formal causal inference, this prior has received little to no attention in the analysis of binary experiments.[3] Our analysis shows that the generalized Dirichlet distribution emerges naturally from the causal formulation of the problem, that the parameters of the distribution can be cast in intuitive clinical terms, and that statistical inference is manageable, with exact posterior sampling and analytical formulae for Bayes factors, which we derive in this paper.[4]

Outline of the chapter. Section 4.2 introduces the statistical setup for the analysis of binary experiments and reviews existing methods for Bayesian inference in this setting. Section 4.3 introduces our proposal. It also derives key results for implementation, such as analytical formulae for the marginal likelihood and algorithms for exact posterior sampling. Section 4.4 demonstrates the utility of our method in three empirical examples. Section 4.5 concludes the chapter, and suggests possible extensions for future research. Code to replicate our analysis is available at https://github.com/njirons/causally-sound.

---

[2]Similar priors have also appeared in the analysis of diagnostic testing, such as in Branscum et al. (2005). This literature seems to be unaware of its connections with the generalized Dirichlet distribution, and some of the results we provide here, such as exact sampling, and analytical formulae for the marginal likelihood, could also be potentially applied to such settings (we leave this to future work).

[3]Related to our setup are studies that have used a *traditional* Dirichlet distribution on response types. This can be shown to be a special case of our proposal, and we discuss it in Sections 4.2.3 and 4.3.

[4]The history of statistical analysis of contingency tables is extensive; Killion and Zahn (1976) and Agresti and Hitchcock (2005) provide comprehensive reviews. Along the lines of relevant studies already mentioned, Tian et al. (2003) and Ng et al. (2008), identify special cases of Dickey's generalized Dirichlet which admit alternative stochastic representations and simplified computation of posterior quantities. Less relevant to our proposed methodology, other priors used to model contingency table proportions have been proposed in Albert and Gupta (1982, 1983a, 1983b, 1985), Basu and Pereira (1982), Leonard (1972, 1975), and Park and Brown (1994).

4.2 Preliminaries

In this section we set notation, the statistical setup, and briefly review the two main approaches currently used for the Bayesian analysis of binary experiments—the independent beta and logit transformation approaches. We also briefly introduce the response type parameterization of the joint distribution of potential outcomes, which is an important stepping stone for understanding our proposal.

4.2.1 Potential outcomes

Our analysis is situated within the potential outcomes framework of causal inference (Neyman, 1990; Rubin, 1974). Let N denote the total number of participants in the study, Z_i a binary treatment indicator and Y_i a binary outcome indicator for subject $i \in \{1, \ldots, N\}$. We denote by $Y_i(z)$ the potential outcome of subject i under the experimental condition $Z_i = z$, where $z = 0$ indicates the control and $z = 1$ the treatment condition. Under the standard consistency assumption, we have that the observed outcome of subject i equals the potential outcome associated to the experimental condition that subject i has actually received, i.e., $Y_i = Y_i(Z_i)$. Throughout the paper, we adopt the convention that $Y_i = 1$ denotes an adverse outcome, such as death or the contraction of a disease. We take a super-population perspective, and assume that subjects are independent and identically distributed (i.i.d.) draws from a common population. We assume complete randomization, which implies ignorability of the treatment assignment, $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp Z_i$.

4.2.2 Marginal parameterization

When subjects are independently drawn from a common super-population and the treatment is assigned at random, it follows that the observed *counts* of adverse outcomes in each

treatment arm,

$$y_0 = \sum_{i=1}^{N} Y_i(1 - Z_i), \qquad y_1 = \sum_{i=1}^{N} Y_i Z_i,$$

follow independent binomial distributions:

$$y_0 \sim \text{Binomial}(N_0, \theta_0) \quad \perp\!\!\!\perp \quad y_1 \sim \text{Binomial}(N_1, \theta_1),$$

where here, $\theta_1 = \mathbb{P}(Y_i(1) = 1)$, $N_1 = \sum_i Z_i$ denote the probability of an adverse outcome and the sample size of the treatment group, and $\theta_0 = \mathbb{P}(Y_i(0) = 1)$, $N_0 = N - N_1$ are the analogous quantities for the control group.[5] We refer to the probabilities θ_0 and θ_1 as the *baseline risk* and *risk of treatment*, respectively.

This defines the likelihood under the marginal parameterization of a binary experiment—so called because the parameters (θ_0, θ_1) are defined in terms of the marginal distribution of the potential outcomes $Y_i(0)$ and $Y_i(1)$:

$$L(\mathcal{D}|\theta_0, \theta_1) = \binom{N_0}{y_0} \theta_0^{y_0} (1 - \theta_0)^{N_0 - y_0} \times \binom{N_1}{y_1} \theta_1^{y_1} (1 - \theta_1)^{N_1 - y_1}, \tag{4.1}$$

where hereafter we denote the observed data by $\mathcal{D} = (y_0, y_1, N_0, N_1)$. To determine the effect of treatment, if any, Bayesian inference is carried out using the posterior distribution of the parameters (θ_0, θ_1), which requires specification of a prior distribution for (θ_0, θ_1). There are two main parameterizations with accompanying priors currently in use, discussed extensively in Agresti and Min (2005) and Dablander et al. (2022)—these are the independent beta (IB) and logit transformation (LT) approaches, which we now discuss.

---

[5]The likelihood of the observed outcomes, conditional on the treatment assignment vector $Z_1, \dots Z_N$, factorizes as $\mathbb{P}(Y_1 \dots, Y_N \mid Z_1 = z_1, \dots, Z_n = z_n) = \mathbb{P}(Y_1(z_1), \dots, Y_N(z_N) \mid Z_1 = z_1, \dots, Z_N = z_N) = \mathbb{P}(Y_1(z_1), \dots, Y_N(z_N)) = \prod_i \mathbb{P}(Y_i(z_i)) = \prod_{i:Z_i=1} \mathbb{P}(Y_i(1)) \prod_{i:Z_i=0} \mathbb{P}(Y_i(0))$, where the first equality is due to consistency, the second equality due to ignorability of the treatment assignment, and the third equality due the assumption that the subjects are i.i.d. draws from a common super-population. Therefore, the data can be seen as a sequence of independent Bernoulli trials, and the counts y_0, y_1 as independent binomials. Note this equivalence does not hold under a finite population perspective; see Ding and Miratrix (2019).

Independent beta (IB) approach

The independent beta (IB) approach (Jeffreys, 1935) assigns the prior[6]

$$\theta_0 \sim \text{Beta}(a_0, b_0) \quad \perp\!\!\!\perp \quad \theta_1 \sim \text{Beta}(a_1, b_1), \tag{4.2}$$

for some hyperparameters $a_0, b_0, a_1, b_1 > 0$. A common specification is $a_0 = b_0 = a_1 = b_1 = 1$, which assigns a uniform distribution to (θ_0, θ_1). This choice of flat priors is usually thought to encode ignorance of (θ_0, θ_1) *a priori*, though it makes strong implicit assumptions as we discuss next. We refer to (4.2) as the IB$(a; b)$ prior, where $a = (a_0, a_1), b = (b_0, b_1)$.[7]

The main advantage of the IB approach is its simplicity. As the beta prior is conjugate to the binomial likelihood, estimation and posterior simulation can be carried out exactly without resorting to approximate sampling algorithms, such as MCMC. Furthermore, marginal likelihoods and Bayes factors, which are widely used for Bayesian hypothesis testing and can be difficult to calculate in general (usually requiring numerical approximation or estimation via posterior simulation), can be calculated analytically (Kass & Raftery, 1995).

A significant drawback of the IB approach is the restrictive assumption of independence between θ_0 and θ_1. In most experimental settings, we would expect our knowledge about the risks in the control and treatment groups to be dependent. For example, if we known that the population prevalence of an infectious disease is approximately 1%, we would expect the prevalence of the disease among those receiving a vaccine to be concentrated around 1% or below, reflecting the common prior belief that it is unlikely that the vaccine would cause the disease. The IB prior fails to accommodate this natural dependence between risks in each arm of the trial. Furthermore, since independence in the prior and the likelihood implies

[6]Here $X \sim \text{Beta}(a, b)$ denotes the probability distribution on the unit interval $[0, 1]$ with Lebesgue density proportional to $x^{a-1}(1-x)^{b-1}$.

[7]Note that if we consider outcomes with multiple categories (e.g, as in Thall et al., 1995), the analogous prior here is to assign independent Dirichlet distributions to the vector of probabilities of each arm of the study. This should not be conflated with assigning a Dirichlet prior to the joint distribution of potential outcomes, which we discuss in Section 4.2.3.

independence *a posteriori*, this failure also extends to the posterior.

Logit Transformation (LT) approach

The logit transformation (LT) approach (Agresti & Hitchcock, 2005; Dablander et al., 2022; Kass & Vaidyanathan, 1992) reparameterizes the model in terms of the logit-transformed risks, by defining the parameters (β, ψ) satisfying

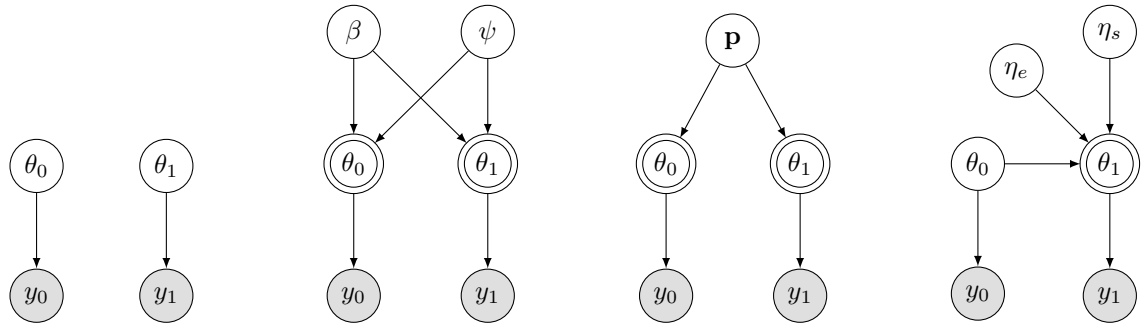
$$\log\left(\frac{\theta_0}{1-\theta_0}\right) = \beta - \frac{\psi}{2}, \qquad \log\left(\frac{\theta_1}{1-\theta_1}\right) = \beta + \frac{\psi}{2}.$$

Note this parameterization is equivalent to a logistic regression of the outcome on the treatment with the encoding $Z \in \{-1/2, 1/2\}$ (Gronau et al., 2021). It then assigns an independent normal prior to (β, ψ):

$$\beta \sim \text{Normal}(\mu_\beta, \sigma^2_\beta) \quad \perp\!\!\!\perp \quad \psi \sim \text{Normal}(\mu_\psi, \sigma^2_\psi), \tag{4.3}$$

where $\mu = (\mu_\beta, \mu_\psi)$ and $\sigma = (\sigma_\beta, \sigma_\psi) > 0$ are hyperparameters. This prior encodes correlation between θ_0 and θ_1 through their shared dependence on β and ψ. We refer to (4.3) as the $\text{LT}(\mu; \sigma)$ prior. Figure 4.1 depicts probabilistic graphical models comparing the IB and LT parameterizations, as well as the other approaches we will introduce in this paper.

While the LT approach induces prior dependence between θ_0 and θ_1, this comes at the cost of a less intuitive parameterization. Here β is interpreted as the "grand log odds," i.e, the average of the log odds across treatment arms, whereas ψ is the log odds ratio. Odds ratios are notoriously difficult to understand, and thus reasoning about the plausible prior means and variances of log odds—two unbounded hyperparameters—is often challenging in practice. The LT approach also has other computational disadvantages relative to the IB prior. Unlike the IB approach, marginal likelihoods and Bayes factors for the LT approach are not available analytically, and posterior sampling must be carried out approximately.



(a) Independent Beta (b) Logit Transform (c) Response Type (d) BREASE

Figure 4.1: Probabilistic graphical models for different parameterizations and prior setups. Gray nodes denote observed variables, white nodes denote latent parameters, and double borders indicate that a node is a deterministic function of its parents. (a) Independent beta priors are placed directly on θ_0 and θ_1; (b) Independent Gaussian priors are placed on the log odds quantities β and ψ; (c) A Dirichlet prior is placed on the response type probabilities $\mathbf{p}$; (d) Our proposal, independent beta priors are placed on θ_0, η_e, and η_s. In all cases, the observed data depends only on θ_0 and θ_1.

4.2.3 Response type (RT) parameterization

The IB and LT approaches focus on the margins of the joint distribution of the potential outcomes $Y_i(0)$ and $Y_i(1)$. This focus is natural, because the observed data depends only upon the parameters θ_0 and θ_1. However, thinking in terms of their *joint* distribution reveals alternative ways of inducing prior dependence between these parameters. Specifically, the joint distribution of potential outcomes is fully characterized by four probabilities

$$p_{jk} = \mathbb{P}(Y_i(0) = j, Y_i(1) = k), \qquad j, k \in \{0, 1\}. \tag{4.4}$$

	$Y_i(0) = 0$	$Y_i(0) = 1$	Row Sum
$Y_i(1) = 0$	$p_{00} = (1 - \eta_s)(1 - \theta_0)$	$p_{10} = \eta_e\theta_0$	$1 - \theta_1$
$Y_i(1) = 1$	$p_{01} = \eta_s(1 - \theta_0)$	$p_{11} = (1 - \eta_e)\theta_0$	θ_1
Column Sum	$1 - \theta_0$	θ_0	

Table 4.1: 2×2 contingency table of potential outcomes for a binary experiment. Only the margins of the table are identified from the observed data.

The probabilities $\mathbf{p} = \{p_{jk}\}_{j,k\in\{0,1\}}$ describe the frequencies of the four possible response types in the population (Copas, 1973; Greenland & Robins, 1986).[8] These include: (i) the "doomed" $\{Y_i(0) = 1, Y_i(1) = 1\}$, for whom the adverse outcome occurs regardless of treatment; (ii) the "immune" $\{Y_i(0) = 0, Y_i(1) = 0\}$, for whom the adverse outcome does not occur regardless of treatment; (iii) the "preventive" $\{Y_i(0) = 1, Y_i(1) = 0\}$, for whom treatment *prevents* the adverse outcome; and, (iv) the "causal" $\{Y_i(0) = 0, Y_i(1) = 1\}$, for whom treatment *causes* the adverse outcome. Here θ_0 and θ_1, which satisfy $\theta_0 = p_{10} + p_{11}$ and $\theta_1 = p_{01} + p_{11}$, define the margins of Table 4.1.

Whereas in the marginal parameterization, independence of the likelihood and prior imply that estimation of θ_0 is only informed by data in the control group (and similarly for θ_1), the response type (RT) parameterization intertwines the data from each arm of the study. The shared dependence of θ_0 and θ_1 on the response type proportions reveals the link between outcomes in the control and treated groups.

A Bayesian approach to modeling the response type probabilities $\mathbf{p}$ requires specification

[8]These probabilities are also known as "probabilities of causation" (Pearl, 2009; Tian & Pearl, 2000); for instance, $\mathbb{P}(Y_i(0) = 1, Y_i(1) = 0)$ is referred by Tian and Pearl (2000) as the probability that the treatment is both necessary and sufficient to prevent an adverse outcome.

of a prior density supported on the probability simplex, making the Dirichlet distribution a natural candidate[9]

$$\mathbf{p} = (p_{00}, p_{10}, p_{01}, p_{11}) \sim \text{Dirichlet}(a_{00}, a_{10}, a_{01}, a_{11}), \qquad a_{00}, a_{10}, a_{01}, a_{11} > 0. \tag{4.5}$$

Indeed, priors of this type have been used in the analysis of partially identified quantities in randomized trials with non-compliance, such as in Chickering and Pearl (1996).[10] As we show next, the Dirichlet prior is a special case of our proposal, and our analysis not only extends it, but also clarifies its advantages and limitations as a means to induce the desired joint prior distribution on the two binomial proportions (θ_0, θ_1).

4.3 The BREASE framework

In this section we introduce the BREASE framework for the analysis of binary experiments. We start by parameterizing the likelihood in terms of the baseline risk, efficacy, and risk of adverse side effects of the treatment. We then propose a jointly independent beta prior distributions on these three parameters, which we show to be a generalization of the Dirichlet prior on the response types. Our proposal has a number of advantages. From a statistical perspective, it induces dependence between the risks in the treatment and control groups, while also enabling exact posterior sampling, and marginal likelihood calculations. From a clinical perspective, this parameterization casts the model in terms of natural quantities appearing frequently in the clinician's vocabulary, thereby facilitating interpretability, elicitation of prior knowledge, and sensitivity analyses.

---

[9]Here $(p_1, \ldots, p_k) \sim \text{Dirichlet}(a_1, \ldots, a_k)$ denotes the probability distribution on the simplex with Lebesgue density proportional to $\prod_{i=1}^{k} p_i^{a_i - 1}$.

[10]See also Hirano et al. (2000), Imbens and Rubin (1997), and Madigan (1999).

4.3.1 Baseline risk, efficacy and adverse side effects

To make things concrete, suppose $Y_i = 1$ denotes death. We define the *efficacy* of the treatment, η_e, as the probability that the treatment *prevents* the death of a patient that would have otherwise died without it:

$$\eta_e = \mathbb{P}(Y_i(1) = 0 | Y_i(0) = 1). \tag{4.6}$$

Similarly, we define the risk of *adverse side effects* of the treatment, η_s, as the probability that the treatment *causes* the death of a patient that would have otherwise been healthy:[11]

$$\eta_s = \mathbb{P}(Y_i(1) = 1 | Y_i(0) = 0). \tag{4.7}$$

These quantities can be interpreted as probabilities of sufficient causation (Cinelli & Pearl, 2021; Tian & Pearl, 2000), i.e., η_e is the probability that treatment is sufficient to save or cure a patient, while η_s is the probability that treatment is sufficient to kill or hurt a patient. They correspond directly to the counterfactual interpretation of what clinicians colloquially refer to as "efficacy" and "side effects" of a drug or vaccine. Indeed, not coincidentally, a commonly used measure in clinical trials called "efficacy", defined as $1 - \theta_1/\theta_0$, equals precisely η_e under the assumption that treatment causes no harm ($\eta_s = 0$).

Applying the law of total probability, we can decompose the risk of treatment in terms of the baseline risk, efficacy, and risk of adverse side effects (BREASE), as

$$\theta_1 = (1 - \eta_e)\theta_0 + \eta_s(1 - \theta_0). \tag{4.8}$$

Table 4.1 shows how the response type probabilities $\mathbf{p}$ can be written as products of θ_0, η_s, and η_e. As with the response type approach, this parameterization highlights the natural

[11] Note these are severe adverse side effects that result in an outcome (e.g, death) opposite to the desired outcome of interest (i.e, survival). In the medical literature, this is sometimes called a "paradoxical reaction" (Smith et al., 2012). Such events could be the result not only of severe adverse biological reactions, but also of other forms of iatrogenesis, such as medical errors.

dependence between θ_0 and θ_1 that is nevertheless easy to miss without framing the problem in the language of potential outcomes. For example, note that θ_0 and θ_1 are functionally independent only under the strong assumption that $\eta_e = 1 - \eta_s$, i.e., the probability of treatment saving a patient is equal to the probability that it doesn't kill one.

Likelihood

Plugging in (4.8), we can rewrite the likelihood (4.1) in terms of $(\theta_0, \eta_e, \eta_s)$.

Theorem 1. *Under (4.1) and (4.6-4.8), the likelihood is*

$$\begin{aligned} L(\mathcal{D}|\theta_0, \eta_e, \eta_s) = &\binom{N_0}{y_0}\binom{N_1}{y_1}\sum_{j=0}^{y_1}\sum_{k=0}^{N_1-y_1}\binom{y_1}{j}\binom{N_1-y_1}{k} \times \theta_0^{y_0+j+k}(1-\theta_0)^{N-(y_0+j+k)} \\ &\times \eta_e^k(1-\eta_e)^j \\ &\times \eta_s^{y_1-j}(1-\eta_s)^{N_1-y_1-k}, \qquad (\theta_0, \eta_e, \eta_s) \in [0,1]^3. \end{aligned} \tag{4.9}$$

Theorem 1 follows from applying the binomial theorem twice. As the likelihood (4.9) is polynomial in $(\theta_0, \eta_e, \eta_s)$, any prior distribution $\pi(\theta_0, \eta_e, \eta_s)$ for which the moments can be explicitly calculated yields an analytical expression for the marginal likelihood. In particular, if

$$\pi(\theta_0, \eta_e, \eta_s) \propto \theta_0^{\alpha_0-1}(1-\theta_0)^{\beta_0-1} \times \eta_e^{\alpha_e-1}(1-\eta_e)^{\beta_e-1} \times \eta_s^{\alpha_s-1}(1-\eta_s)^{\beta_s-1}$$

is a product of independent beta distributions, as we will see in the next section, then the marginal likelihood is a weighted sum of beta function values. Furthermore, the posterior distribution $\pi(\theta_0, \eta_e, \eta_s|\mathcal{D})$ will be a mixture of independent beta distributions, from which we can sample exactly via simulation.

Partial identification and monotonicity

The parameters η_e and η_s are only partially identified by the observed data. That is, without further assumptions, we have the following bounds,

$$\max\left\{0, 1-\frac{\theta_1}{\theta_0}\right\} \leq \eta_e \leq \min\left\{\frac{1-\theta_1}{\theta_0}, 1\right\}, \quad \max\left\{0, \frac{\theta_1-\theta_0}{1-\theta_0}\right\} \leq \eta_s \leq \min\left\{\frac{\theta_1}{1-\theta_0}, 1\right\}.$$

Thus, as the sample size increases, the posterior distribution of η_s and η_e will not concentrate in a point—rather, it will remain spread over its partially identified region (Gustafson, 2015; Richardson et al., 2011). Notice, however, that this does not affect the behavior of the posterior distribution of (θ_0, θ_1). The BREASE parameterization thus explicitly separates the identified and partially identified parameters—(θ_0, θ_1) and (η_e, η_s), respectively. Even if interest does not lie in the counterfactual probabilities (η_s, η_e) *per se*, assigning a prior to those quantities can be thought of as a causally principled way to specify a joint prior on the identified target parameters (θ_0, θ_1).

Finally, a common assumption in the potential outcomes literature is called *monotonicity*, which states that the treatment does no harm. In our framework, this corresponds to the constraint $\eta_s = 0$. This assumption is reasonable in many clinical settings. Under monotonicity, the efficacy of the treatment is in fact point identified, and given by $\eta_e = 1 - \theta_1/\theta_0$. The quantity θ_1/θ_0 is known as the risk ratio, and the quantity $1 - \theta_1/\theta_0$ is indeed known as "efficacy" in the clinical trials literature. While the hard constraint $\eta_s = 0$ may not be credible in some settings, if side effects are expected to be small, the BREASE approach allows one to instead place an informative prior on η_s.

4.3.2 Prior specification

Bayesian inference with the likelihood (4.9) requires specifying a prior distribution on three separate and variation independent probabilities, $(\theta_0, \eta_e, \eta_s)$. We propose setting jointly

independent beta prior distributions on these parameters:

$$\theta_0 \sim \text{Beta}^*(\mu_0, n_0) \quad \perp\!\!\!\perp \quad \eta_e \sim \text{Beta}^*(\mu_e, n_e) \quad \perp\!\!\!\perp \quad \eta_s \sim \text{Beta}^*(\mu_s, n_s), \tag{4.10}$$

where here $\text{Beta}^*(\mu, n)$ denotes a $\text{Beta}(a, b)$ distribution, with mean $\mu = a/(a+b)$ and prior "sample size" $n = a+b$. We refer to (4.10) as the BREASE$(\mu; n)$ prior, where $\mu = (\mu_0, \mu_e, \mu_s)$, $n = (n_0, n_e, n_s)$.

Since (4.10) defines a jointly independent beta prior on $(\theta_0, \eta_e, \eta_s)$, the discussion in Section 4.3.1 applies. In particular, the posterior of $(\theta_0, \eta_e, \eta_s)$ is a mixture of independent betas, which permits exact sampling via simulation, and the marginal likelihood is available analytically as a weighted sum of beta functions, as we show in Sections 4.3.5 and 4.3.6.

Connections to the (generalized) Dirichlet. The prior (4.10) induces a *generalized* Dirichlet distribution (Dickey et al., 1987; Dickey, 1983; Tian et al., 2003) on the vector of potential outcomes probabilities $\mathbf{p}$—see Section 4.3.4 for derivation and further discussion. In particular, the generalized Dirichlet reduces to the traditional Dirichlet distribution (4.5) for the following restricted choice of prior sample sizes

$$n_e = \mu_0 n_0, \quad n_s = (1-\mu_0) n_0. \tag{4.11}$$

Moreover, since $\theta_1 = p_{01} + p_{11}$, by the aggregation property of the Dirichlet (Ng et al., 2011), marginally we have

$$\theta_1 \sim \text{Beta}^* \left((1-\mu_e)\mu_0 + \mu_s(1-\mu_0), n_0 \right), \tag{4.12}$$

which mirrors the decomposition (4.8). The BREASE approach thus reveals an implicit "equal confidence" assumption of the Dirichlet: the prior spread for θ_0 determines the spread of the distributions of η_e, η_s, and θ_1 *a priori*. Hence, the Dirichlet is underparameterized, and unsuitable for cases in which, say, we have ample knowledge of the baseline risk but relatively little information about the possible efficacy or side effects of the treatment (or

vice-versa). Casting the likelihood in terms of the BREASE parameters makes such choices explicit, by allowing the hyperparameters governing θ_0, η_e and η_s to be set independently.

Induced prior distribution of (θ_0, θ_1)

As mentioned in Section 4.3.1, our goal with the BREASE approach is primarily to induce causally sound priors on the identified parameters of interest, the two binomial proportions (θ_0, θ_1). Thus we now discuss the induced marginal and conditional distribution of the risk of treatment, θ_1, under the BREASE prior (4.10).

From equation (4.8) we see that θ_1, conditionally on θ_0, is distributed as a convex combination of independent beta random variables *a priori*. This distribution was studied in Pham-Gia and Turkkan (1998) and is given in terms of Appell's first hypergeometric function F_1—in Section 4.3.3 we derive the explicit formula and provide further discussion. From here, the marginal prior on θ_1 can be obtained as $\pi(\theta_1) = \int_0^1 \pi(\theta_1|\theta_0)\pi(\theta_0)d\theta_0$. While the general formula for $\pi(\theta_1|\theta_0)$ may look unwieldy, and the integration in $\pi(\theta_1)$ prohibitive, there are noteworthy specific cases.

Equal confidence. As noted in the previous discussion, under the equal confidence assumption, $n_e = \mu_0 n_0$, $n_s = (1-\mu_0)n_0$, the marginal prior induced on θ_1 is the beta distribution in (4.12). In particular, to obtain equal marginal priors for the treatment and control groups, i.e., $\theta_z \sim \text{Beta}(\mu_0, n_0)$ for $z \in \{0, 1\}$, it suffices to set $\mu_s = (\mu_0/(1-\mu_0))\mu_e$, with $0 \leq \mu_e \leq \min(1, (1-\mu_0)/\mu_0)$. Choosing $\mu_0 = 1/2$, $n_0 = 2$, and $\mu_e = \mu_s = \mu$ results in marginal uniform priors with prior correlation $\text{Cor}(\theta_0, \theta_1) = 1 - 2\mu$.

Uniform prior. When at least one of η_e, η_s is uniformly distributed, the conditional prior $\pi(\theta_1|\theta_0)$ reduces to a simple expression in terms of the CDF of the beta distribution, which we derive in Section 4.3.3. In particular, with a flat prior $(\theta_0, \eta_e, \eta_s) \sim \text{Uniform}(0,1)^3$, the marginal on θ_1 is $\pi(\theta_1) = -2\theta_1 \log \theta_1 - 2(1-\theta_1)\log(1-\theta_1)$.

Moments. The joint density $\pi(\theta_0, \theta_1)$ induced by the BREASE$(\mu; n)$ prior is generally complicated, but its moments are easily computed in terms of the hyperparameters (μ, n) as θ_1 is a polynomial in $(\theta_0, \eta_e, \eta_s)$, which are beta distributed *a priori*. For example, the prior covariance has a simple form, $\text{Cov}(\theta_0, \theta_1) = \frac{\mu_0(1-\mu_0)}{n_0+1}(1 - \mu_e - \mu_s)$. This implies the following directions of the prior correlation,

$$\text{Cor}(\theta_0, \theta_1) \begin{cases} < 0, & \mu_e + \mu_s > 1, \\ = 0, & \mu_e + \mu_s = 1, \\ > 0, & \mu_e + \mu_s < 1. \end{cases} \tag{4.13}$$

In words, θ_0 and θ_1 are positively correlated *a priori* when the expected harm and benefit of treatment are small, and negatively correlated otherwise.

Default prior. While we encourage the use of informative priors, it is useful to have reasonable defaults to start the analysis. If we would like to put θ_0 and θ_1 on equal footing, the BREASE$(1/2, \mu, \mu; 2, 1, 1)$ is thus the natural choice, with the following properties: (i) puts flat uniform priors on θ_0 and θ_1 (as with the IB approach); (ii) induces prior correlation between parameters (as with the LT approach); (iii) assumes no effect of treatment, on average (as with the IB and LT approaches); and, (iv) depends on a single, easily interpretable parameter μ denoting the expected benefits (efficacy) or harm (side effects) of the treatment. When $\mu > 1/2$, θ_1 and θ_0 become anti-correlated, and thus for most cases, $\mu \leq 1/2$ is a reasonable choice. Our preferred specification uses $\mu = 0.3$ as the default. As Figure 4.2 shows, this (weakly) encodes the expectation of moderate effects and concentrates mass on the diagonal $\theta_0 = \theta_1$. This quality is useful in the context of Bayesian hypothesis testing. When testing a null hypothesis H_0 (e.g., no effect of treatment on average, $H_0 : \theta_0 = \theta_1$) nested within an alternative H_1, it is desirable for the prior under H_1 to concentrate mass around the null model (Casella & Moreno, 2009; Gunel & Dickey, 1974; Jeffreys, 1961).

Comparison to the LT and IB priors. When comparing estimates and inferences resulting from the IB, LT, and BREASE priors, one may wonder how much of the differences are driven by the induced marginal priors on θ_0 and θ_1 and how much are due to the prior correlation between θ_0 and θ_1. To address this concern, we introduce two versions of the LT prior calibrated to match the default BREASE prior in certain senses by varying the dispersion hyperparameters σ_β and σ_ψ. Figure 4.3 exhibits these calibrated LT priors (as well as the default LT with $(\sigma_\beta, \sigma_\psi) = (1, 1)$ in the middle column).

The first specification $(\sigma_\beta, \sigma_\psi) = (1.3, 1.67)$ in the leftmost column places near-uniform marginals on θ_0 and θ_1 and induces a prior correlation between θ_0 and θ_1 approximating 0.4, as with the default BREASE prior. This choice of hyperparameters does not concentrate mass on the diagonal $\theta_0 = \theta_1$, and, as we will see in Section 4.4, the resulting Bayes factors may differ from those of the default BREASE prior.

The second specification $(\sigma_\beta, \sigma_\psi) = (1.5, 0.1)$ in the rightmost column was visually calibrated to place near-uniform marginals on θ_0 and θ_1 and concentrate mass on the diagonal $\theta_0 = \theta_1$, as with the default BREASE. Comparing Figures 4.2 and 4.3, we might expect these two priors to behave similarly. Nevertheless, this choice of hyperparameters induces prior correlation $\text{Cor}(\theta_0, \theta_1) \approx 0.998$. This is more informative than the default BREASE prior (which has correlation 0.4 and uses unit information priors on η_e and η_s). As a result, this LT prior can lead to a substantial amount of shrinkage, as we demonstrate in Section 4.4.

From these observations and the results in Section 4.4, we conclude that the LT prior may not be flexible enough to encode BREASE beliefs: it can weakly encode correlation between θ_0 and θ_1 or concentrate mass on the diagonal, but not both.

Finally, we note that the only IB prior with uniform marginals is precisely the joint independent uniform prior on (θ_0, θ_1), which has $\text{Cor}(\theta_0, \theta_1) = 0$ and will form the basis for our comparison of the IB approach to BREASE in Section 4.4. The differences in the resulting Bayes factors noted therein, which are substantial in the aspirin example, must be

due to the prior correlation rather than the marginal specification.

4.3.3 Implied prior on θ_1

Let the prior of $(\theta_0, \eta_e, \eta_s)$ consist of independent beta distributions with PDFs denoted by $\theta_0 \sim \pi_{\theta_0}(\theta_0)$, $\eta_s \sim \pi_s(\eta_s)$, and $\eta_e \sim \pi_e(\eta_e)$. By the law of total probability, the conditional distribution of θ_1 given θ_0 can be written as

$$\pi(\theta_1 \mid \theta_0) = \int_0^1 \pi(\theta_1 \mid \theta_0, \eta_e)\pi_e(\eta_e)d\eta_e, \tag{4.14}$$

where here we make use of the fact that η_e and θ_0 are *a priori* independent. Note that, conditional on θ_0 and η_e, θ_1 is simply a linear transformation of η_s, namely $\theta_1 = \theta_0(1-\eta_e) + (1-\theta_0)\eta_s$. We can thus write the density of θ_1 in terms of the density of η_s as

$$\pi(\theta_1 \mid \theta_0, \eta_e) = \left(\frac{1}{1-\theta_0}\right) \pi_s \left(\frac{\theta_1 - \theta_0(1-\eta_e)}{1-\theta_0}\right),$$

where we make use of the fact that $\dfrac{d\eta_s}{d\theta_1} = \dfrac{1}{1-\theta_0}$. Substituting this back into Eq. 4.14, we have the following integral

$$\pi(\theta_1 \mid \theta_0) = \left(\frac{1}{1-\theta_0}\right) \int_0^1 \pi_s \left(\frac{\theta_1 - \theta_0(1-\eta_e)}{1-\theta_0}\right) \pi_e(\eta_e)d\eta_e. \tag{4.15}$$

For the special case where η_e is uniformly distributed, $\pi_e(\eta_e) = 1$, the integral simplifies,

$$\pi(\theta_1 \mid \theta_0) = \left(\frac{1}{1-\theta_0}\right) \int_0^1 \pi_s \left(\frac{\theta_1 - \theta_0(1-\eta_e)}{1-\theta_0}\right) d\eta_e \tag{4.16}$$

$$= \left(\frac{1}{\theta_0}\right) \int_{\frac{\theta_1-\theta_0}{1-\theta_0}}^{\frac{\theta_1}{1-\theta_0}} \pi_s\left(\eta_s\right) d\eta_s \tag{4.17}$$

$$= \left(\frac{1}{\theta_0}\right) \left(F_s\left(\frac{\theta_1}{1-\theta_0}\right) - F_s\left(\frac{\theta_1-\theta_0}{1-\theta_0}\right)\right), \tag{4.18}$$

where the second equality follows from change of variables, noting $d\eta_e = (1-\theta_0)/\theta_0 d\eta_s$. Here $F_s(\cdot)$ denotes the CDF of the beta distribution, which is given by the the regularized incomplete beta function.

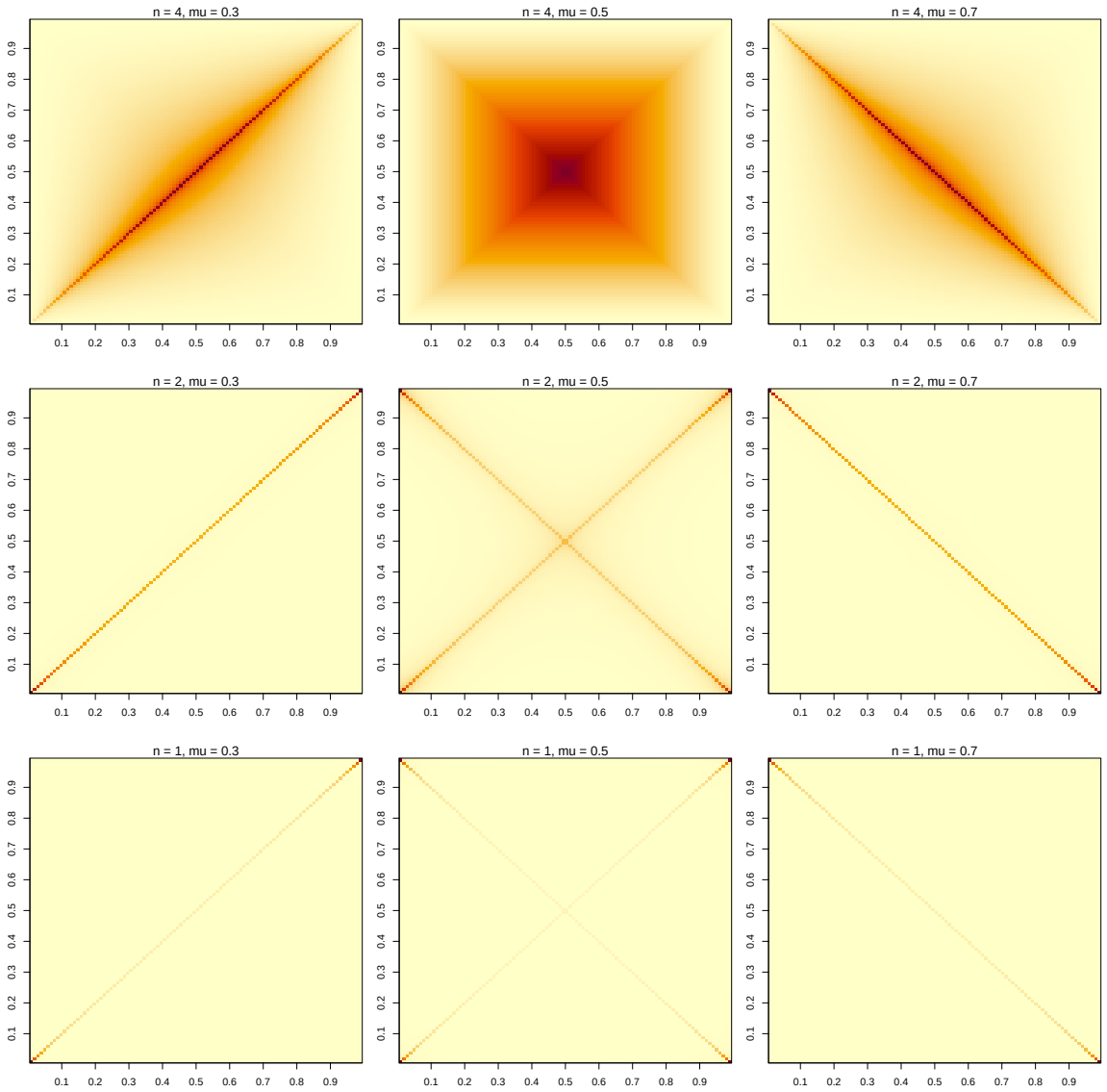


Figure 4.2: Heatmaps of the joint density of (θ_0, θ_1) under the BREASE$(1/2, \mu, \mu; n, n/2, n/2)$ prior varying n and μ. Our proposed default prior takes $n = 2$ and $\mu = .3$. As the plot shows, this: (i) leads to uniform marginals on θ_0 and θ_1; (ii) assumes zero treatment effect on average; (iii) concentrates mass on the diagonal $\theta_0 = \theta_1$; (iv) favors small (or large) proportions, instead of proportions around the center, which is expected when one quantifies rare outcomes such as death (proportions would be small) or, its complement, survival (in which case proportions would be large).

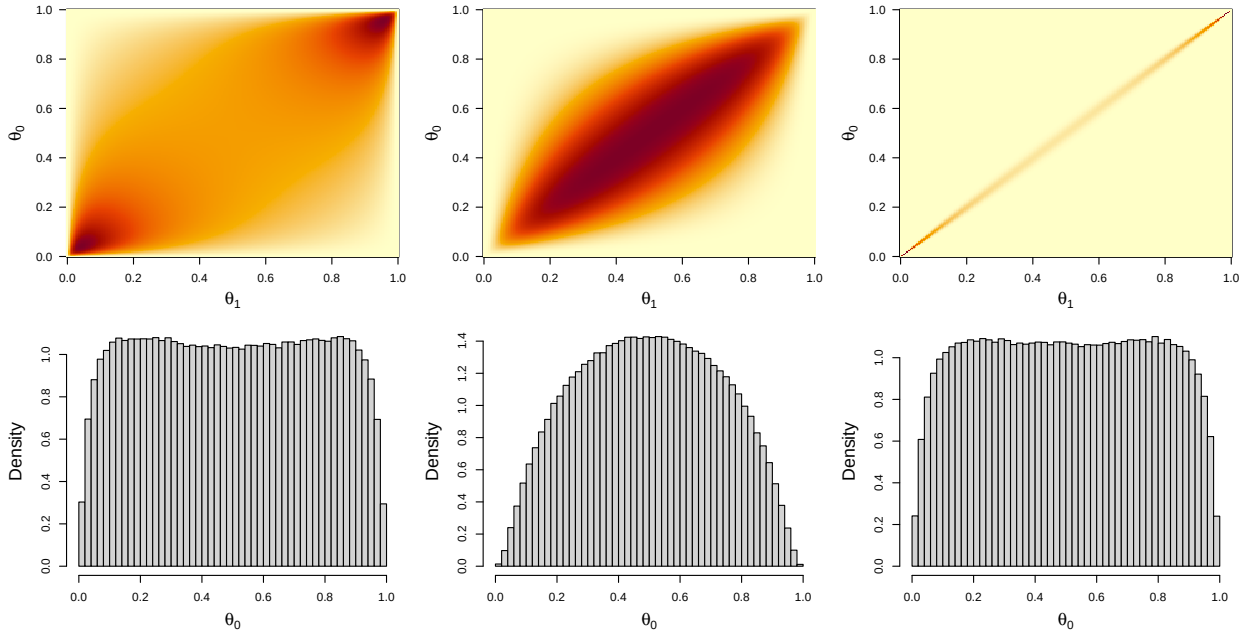


Figure 4.3: **Top row:** heatmaps of the joint density of (θ_0, θ_1) under the LT prior varying σ_β and σ_ψ: (i) $(\sigma_\beta, \sigma_\psi) = (1.3, 1.67)$, calibrated to put near-uniform marginals on θ_0 and θ_1 and prior correlation $\mathrm{Cor}(\theta_0, \theta_1) \approx 0.4$ matching the default BREASE prior; (ii) $(\sigma_\beta, \sigma_\psi) = (1, 1)$, the default LT prior with correlation $\mathrm{Cor}(\theta_0, \theta_1) \approx 0.59$; (iii) $(\sigma_\beta, \sigma_\psi) = (1.5, 0.1)$, (visually) calibrated to put near-uniform marginals on θ_0 and θ_1 and concentrate on the diagonal $\theta_0 = \theta_1$, with prior correlation $\mathrm{Cor}(\theta_0, \theta_1) \approx 0.998$. **Bottom row:** histograms of the marginal densities of θ_0 (and, by symmetry, θ_1) under each of the LT priors above.

For special cases the expression above simplifies. For instance, when η_s is also uniformly distributed, we have that $F_s(x) = x$, and we obtain a simple closed form expression for the

conditional density. Specifically, for $\theta_0 \leq 1/2$,

$$\pi(\theta_1 \mid \theta_0) = \begin{cases} \dfrac{\theta_1}{\theta_0(1-\theta_0)} & \text{if } 0 \leq \theta_1 < \theta_0, \\ \dfrac{1}{1-\theta_0} & \text{if } \theta_0 \leq \theta_1 < 1-\theta_0, \\ \dfrac{1-\theta_1}{\theta_0(1-\theta_0)} & \text{if } 1-\theta_0 \leq \theta_1 \leq 1, \end{cases} \tag{4.19}$$

and zero, otherwise. Analogously, for $\theta_0 \geq 1/2$,

$$\pi(\theta_1 \mid \theta_0) = \begin{cases} \dfrac{\theta_1}{\theta_0(1-\theta_0)} & \text{if } 0 \leq \theta_1 < 1-\theta_0, \\ \dfrac{1}{\theta_0} & \text{if } 1-\theta_0 \leq \theta_1 < \theta_0, \\ \dfrac{1-\theta_1}{\theta_0(1-\theta_0)} & \text{if } \theta_0 \leq \theta_1 \leq 1, \end{cases} \tag{4.20}$$

and zero, otherwise. Notice this is a piece-wise linear function of θ_1. Remarkably, however, integrating each region over θ_0 results in the following marginal distribution of $\pi(\theta_1)$,

$$\pi(\theta_1) = 2(-\theta_1 \log \theta_1 - (1-\theta_1)\log(1-\theta_1)),$$

for $\theta_1 \in [0,1]$, and zero otherwise, which is twice the entropy of the Bernoulli(θ_1) distribution.

More generally, the distribution of linear combinations of beta random variables was studied in Pham-Gia and Turkkan (1998) and is given in terms of Appell's first hypergeometric function F_1, which is an infinite series in two variables:

$$F_1(x, y; a; b_1, b_2; c) = \sum_{m_1=0}^{\infty} \sum_{m_2=0}^{\infty} \frac{\Gamma(a+m_1+m_2)\Gamma(b_1+m_1)\Gamma(b_2+m_2)\Gamma(c)}{\Gamma(a)\Gamma(b_1)\Gamma(b_2)\Gamma(c+m_1+m_2)} \frac{x^{m_1}}{m_1!} \frac{y^{m_2}}{m_2!}. \tag{4.21}$$

Appell's function also has an integral representation given by

$$F_1(x, y; a; b_1, b_2; c) = \mathrm{B}(a, c-a)^{-1} \int_0^1 u^{a-1}(1-u)^{c-a-1}(1-ux)^{-b_1}(1-uy)^{-b_2} du. \tag{4.22}$$

Applying the results of Pham-Gia and Turkkan (1998) to our setup, the prior on θ_1 conditional on θ_0 induced by the BREASE prior can be obtained as the following piecewise function: (i) for $\theta_0 \leq 1/2$, we have

$$
\begin{aligned}
\pi(\theta_1|\theta_0) = {} & I(0 \leq \theta_1 \leq \theta_0) \\
& \times \frac{\theta_1^{(1-\mu_e)n_e+\mu_s n_s-1}(\theta_0-\theta_1)^{\mu_e n_e-1}\mathrm{B}(\mu_s n_s,(1-\mu_e)n_e)}{\theta_0^{n_e-1}(1-\theta_0)^{\mu_s n_s}\mathrm{B}(\mu_s n_s,(1-\mu_s)n_s)\mathrm{B}((1-\mu_e)n_e,\mu_e n_e)} \\
& \times F_1\left(\frac{-\theta_1}{\theta_0-\theta_1},\frac{\theta_1}{1-\theta_0};\mu_s n_s;1-\mu_e n_e,1-(1-\mu_s)n_s;(1-\mu_e)n_e+\mu_s n_s\right) \\
& + I(\theta_0 \leq \theta_1 \leq 1-\theta_0) \\
& \times \frac{(\theta_1-\theta_0)^{\mu_s n_s-1}(1-\theta_1)^{(1-\mu_s)n_s-1}}{(1-\theta_0)^{n_s-1}\mathrm{B}(\mu_s n_s,(1-\mu_s)n_s)} \\
& \times F_1\left(\frac{-\theta_0}{\theta_1-\theta_0},\frac{\theta_0}{1-\theta_1};\mu_e n_e;1-\mu_s n_s,1-(1-\mu_s)n_s;n_e\right) \\
& + I(1-\theta_0 \leq \theta_1 \leq 1) \\
& \times \frac{(1-\theta_1)^{\mu_e n_e+(1-\mu_s)n_s-1}(\theta_1-\theta_0)^{\mu_s n_s-1}\mathrm{B}(\mu_e n_e,(1-\mu_s)n_s)}{\theta_0^{\mu_e n_e}(1-\theta_0)^{n_s-1}\mathrm{B}(\mu_s n_s,(1-\mu_s)n_s)\mathrm{B}((1-\mu_e)n_e,\mu_e n_e)} \\
& \times F_1\left(\frac{1-\theta_1}{\theta_0},\frac{\theta_1-1}{\theta_1-\theta_0};\mu_e n_e;1-(1-\mu_e)n_e,1-\mu_s n_s;\mu_e n_e+(1-\mu_s)n_s\right).
\end{aligned}
\tag{4.23}
$$

Similarly, (ii) for $\theta_0 \geq 1/2$, we have

$$\begin{aligned}
\pi(\theta_1|\theta_0) = {} & I(0 \leq \theta_1 \leq 1-\theta_0) \\
& \times \frac{\theta_1^{(1-\mu_e)n_e+\mu_s n_s-1}(1-\theta_0-\theta_1)^{(1-\mu_s)n_s-1}\mathrm{B}((1-\mu_e)n_e,\mu_s n_s)}{(1-\theta_0)^{n_s-1}\theta_0^{(1-\mu_e)n_e}\mathrm{B}((1-\mu_e)n_e,\mu_e n_e)\mathrm{B}(\mu_s n_s,(1-\mu_s)n_s)} \\
& \times F_1\left(\frac{-\theta_1}{1-\theta_0-\theta_1},\frac{\theta_1}{\theta_0};(1-\mu_e)n_e;1-(1-\mu_s)n_s,1-\mu_e n_e;(1-\mu_e)n_e+\mu_s n_s\right) \\
& + I(1-\theta_0 \leq \theta_1 \leq \theta_0) \\
& \times \frac{(\theta_1-(1-\theta_0))^{(1-\mu_e)n_e-1}(1-\theta_1)^{\mu_e n_e-1}}{\theta_0^{n_e-1}\mathrm{B}((1-\mu_e)n_e,\mu_e n_e)} \\
& \times F_1\left(\frac{-(1-\theta_0)}{\theta_1-(1-\theta_0)},\frac{1-\theta_0}{1-\theta_1};(1-\mu_s)n_s;1-(1-\mu_e)n_e,1-\mu_e n_e;n_s\right) \\
& + I(\theta_0 \leq \theta_1 \leq 1) \\
& \times \frac{(1-\theta_1)^{\mu_e n_e+(1-\mu_s)n_s-1}(\theta_1-(1-\theta_0))^{(1-\mu_e)n_e-1}\mathrm{B}((1-\mu_s)n_s,\mu_e n_e)}{(1-\theta_0)^{(1-\mu_s)n_s}\theta_0^{n_e-1}\mathrm{B}((1-\mu_e)n_e,\mu_e n_e)\mathrm{B}(\mu_s n_s,(1-\mu_s)n_s)} \\
& \times F_1\left(\frac{1-\theta_1}{1-\theta_0},\frac{\theta_1-1}{\theta_1-(1-\theta_0)};(1-\mu_s)n_s;1-\mu_s n_s,1-(1-\mu_e)n_e;\mu_e n_e+(1-\mu_s)n_s\right).
\end{aligned} \tag{4.24}$$

Monotonicity. Under the "no harm" monotonicity assumption $\eta_s = 0$ we have $\theta_1 = (1-\eta_e)\theta_0$, in which case θ_1 is a product of independent beta random variables *a priori*. Springer and Thompson (1970) derived the form of this distribution, with the density given as a Meijer G-function. In general, this function is expressed as a contour integral in the complex plane. However, when $a_e = \mu_e n_e$, $b_e = (1-\mu_e)n_e$, $a_0 = \mu_0 n_0$, and $b_0 = (1-\mu_0)n_0$ are integers, the prior on θ_1 can be expressed in closed form as

$$\pi(\theta_1) = \frac{\Gamma(n_0)\Gamma(n_e)}{\Gamma(\mu_0 n_0)\Gamma((1-\mu_e)n_e)} \sum_{k=1}^{m}\sum_{j=0}^{e_k-1} \frac{K_{kj}\theta_1^{d_k-1}(-\log\theta_1)^{e_k-j-1}}{\Gamma(e_k-j)\Gamma(j+1)},$$

where $\{d_1, \ldots, d_m\}$ denote the m different integers occurring with multiplicity $\{e_1, \ldots, e_m\}$, respectively, among the sets $\{a_0 - 1, \ldots, a_0 + b_0 - 2\}$ and $\{a_e - 1, \ldots, a_e + b_e - 2\}$, and

$$K_{kj} = \sum_{r=0}^{j} \sum_{q \in \{1,\ldots,m\}, q \neq k} (-1)^{r+1} \binom{j}{r} \frac{\Gamma(r+1)e_q}{(d_q - d_k)^{r+1}}.$$

In particular, if $a_e + b_e = a_0$ (equivalently $n_e = \mu_0 n_0$, an implicit assumption of the Dirichlet prior), we have

$$\theta_1 \sim \text{Beta}((1-\mu_e)n_e, \mu_e n_e + (1-\mu_0)n_0).$$

For another example, if $(\theta_0, \eta_e) \sim \text{Uniform}(0,1)^2$, we have

$$\pi(\theta_1) = -\log \theta_1.$$

Regarding the conditional prior $\pi(\theta_1|\theta_0)$ under the "no harm" assumption, it is clearly a scaled beta distribution, since $\theta_1 = (1-\eta_e)\theta_0$. If $\eta_e \sim \text{Uniform}(0,1)$, we then have that $\theta_1|\theta_0 \sim \text{Uniform}(0, \theta_0)$. Similarly, under the "no benefit" assumption $\eta_e = 0$, we have that $\theta_1 = \theta_0 + \eta_s(1-\theta_0)$, which is a scaled and shifted beta random variable conditional on θ_0. If $\eta_s \sim \text{Uniform}(0,1)$, then $\theta_1|\theta_0 \sim \text{Uniform}(\theta_0, 1)$.

As for the moments, applying the law of total covariance to the terms involving θ_1 by conditioning on θ_0 and making use of equation (4.8), we obtain

$$\begin{aligned}
\text{Cov}(\theta_0, \theta_1) &= \frac{\mu_0(1-\mu_0)}{n_0+1}(1-\mu_e-\mu_s), \\
\text{Var}(\theta_0) &= \frac{\mu_0(1-\mu_0)}{n_0+1}, \\
\text{Var}(\theta_1) &= \frac{\mu_0(1-\mu_0)}{n_0+1}(1-\mu_e-\mu_s)^2 \\
&\quad + \frac{\mu_e(1-\mu_e)}{n_e+1}\left\{\frac{\mu_0(1-\mu_0)}{n_0+1} + \mu_0^2\right\} \\
&\quad + \frac{\mu_s(1-\mu_s)}{n_s+1}\left\{\frac{\mu_0(1-\mu_0)}{n_0+1} + (1-\mu_0)^2\right\}.
\end{aligned}$$

This can be used to obtain the prior correlation,

$$\text{Cor}(\theta_0, \theta_1) = \frac{\text{Cov}(\theta_0, \theta_1)}{\sqrt{\text{Var}(\theta_0)\text{Var}(\theta_1)}}.$$

4.3.4 The generalized Dirichlet distribution on $\mathbf{p}$

Given a vector of probabilities $\mathbf{p} = (p_1, \ldots, p_k)$, such that $\sum_{i=1}^{k} p_i = 1$, the generalized Dirichlet distribution (Tian et al., 2003) is defined as,

$$\pi(\mathbf{p}) \propto \prod_{i=1}^{k} p_i^{a_i - 1} \prod_{j=1}^{m} \left(\sum_{i=1}^{k} \gamma_{ij} p_i \right)^{b_j - 1} \tag{4.25}$$

where $\Gamma = (\gamma_{ij})$ is a $k \times m$ known scale matrix. We refer to (4.25) as $\mathrm{GD}(a, b, \Gamma)$. Now consider the vector of potential outcomes $\mathbf{p} = (p_{00}, p_{01}, p_{10}, p_{11})$. By change of variables arguments, if $(\theta_0, \eta_e, \eta_s) \sim \mathrm{BREASE}(\mu; n)$ as in (4.10), it is easy to show that $\mathbf{p}$ has density

$$\pi(\mathbf{p}) \propto p_{00}^{(1-\mu_s)n_s - 1} p_{01}^{\mu_s n_s - 1} p_{10}^{\mu_e n_e - 1} p_{11}^{(1-\mu_e)n_e - 1} (p_{00} + p_{01})^{(1-\mu_0)n_0 - n_s} (p_{10} + p_{11})^{\mu_0 n_0 - n_e}, \tag{4.26}$$

which is a $\mathrm{GD}(a, b, \Gamma)$ distribution with parameters

$$\begin{aligned} a &= (\mu_s n_s, (1-\mu_s)n_s, \mu_e n_e, (1-\mu_e)n_e), \\ b &= ((1-\mu_0)n_0 - n_s + 1, \mu_0 n_0 - n_e + 1, 1, 1), \\ \Gamma &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \end{aligned}$$

The prior (4.26) is also a grouped Dirichlet distribution, as defined in Tian et al. (2003) and Ng et al. (2008) (which is a special case of the generalized Dirichlet). Similarly, the posterior in (4.30) induces the following posterior distribution on the vector $\mathbf{p}$,

$$\begin{aligned} \pi(\mathbf{p}|\mathcal{D}) \propto\ & p_{00}^{(1-\mu_s)n_s - 1} p_{01}^{\mu_s n_s - 1} p_{10}^{\mu_e n_e - 1} p_{11}^{(1-\mu_e)n_e - 1} \\ & \times (p_{00} + p_{01})^{N_0 - y_0 + (1-\mu_0)n_0 - n_s} (p_{10} + p_{11})^{y_0 + \mu_0 n_0 - n_e} \\ & \times (p_{00} + p_{10})^{N_1 - y_1} (p_{01} + p_{11})^{y_1}, \end{aligned}$$

which is again a generalized Dirichlet distribution, $\text{GD}(a, b', \Gamma)$, with parameters a and Γ as in the prior, and updated parameter b' given by

$$b' = (N_0 + y_0 + (1 - \mu_0)n_0 - n_s + 1, y_0 + \mu_0 n_0 - n_e + 1, N_1 - y_1 + 1, y_1 + 1).$$

The generalized Dirichlet distribution of Dickey (1983), as well as special cases, such as the grouped Dirichlet and Dirichlet-beta, have been proposed for the Bayesian analysis of categorical data and contingency tables with missing observations (Antelman, 1972; Dickey et al., 1987; Gunel, 1984; Karson & Wrobleski, 1970; Kaufman & King, 1973; Ng et al., 2008; Tian et al., 2003). These studies largely focused on the derivation of closed-form expressions (when available) and accurate approximations for posterior moments and predictive probabilities used in estimation and inference. They did not address the parameterization and interpretation of the generalized Dirichlet in terms of the baseline risk, efficacy, and side effects; algorithms for exact posterior simulation; testing for an effect of treatment and sensitivity analysis using analytical formulae; or the specific application to and prior elicitation for binary experiments.

The Dirichlet as a product of independent betas. To better understand the connection of the BREASE prior with the traditional Dirichlet distribution, it is instructive to first derive the distribution of $(\theta_0, \eta_e, \eta_s)$ induced by a Dirichlet prior on the response type probabilities $\mathbf{p}$. The BREASE parameters can be expressed as

$$\theta_0 = p_{10} + p_{11}, \qquad \eta_e = \frac{p_{10}}{p_{10} + p_{11}}, \qquad \eta_s = \frac{p_{01}}{p_{00} + p_{01}}.$$

Elementary properties of the Dirichlet distribution then imply that these quantities are mutually independent beta random variables (Ng et al., 2011)

$$\theta_0 \sim \text{Beta}(a_{10} + a_{11}, a_{00} + a_{01}) \quad \perp\!\!\!\perp \quad \eta_e \sim \text{Beta}(a_{10}, a_{11}) \quad \perp\!\!\!\perp \quad \eta_s \sim \text{Beta}(a_{01}, a_{00}). \quad (4.27)$$

Similarly, since $\theta_1 = p_{01} + p_{11}$, we also have that $\theta_1 \sim \text{Beta}(a_{01} + a_{11}, a_{00} + a_{10})$ marginally.

While the Dirichlet density seems like a natural choice for the probability vector $\mathbf{p}$, the implied distribution on $(\theta_0, \eta_e, \eta_s)$ reveals some implicit assumptions. In particular, this prior has the peculiar (and potentially undesirable) feature that once we have decided on the parameters underlying the marginal distribution of the efficacy and side effects of treatment (η_e, η_s)—which requires specifying $(a_{00}, a_{10}, a_{01}, a_{11})$—we have fully determined the joint prior on $(\theta_0, \eta_e, \eta_s)$. In this sense, the Dirichlet distribution is underparametrized.

This underparameterization becomes clearer with an alternative representation of the beta distribution, in terms of the prior mean and prior "sample size." For $\mu = a/(a+b)$ and $n = a+b$, we write $\text{Beta}^*(\mu, n)$ to denote a $\text{Beta}(a, b)$ distribution, with mean μ and sample size n. The Dirichlet joint prior on $(\theta_0, \eta_e, \eta_s)$ has then the following alternative stochastic representation,

$$\theta_0 \sim \text{Beta}^*(\mu_0, n_0) \quad \perp\!\!\!\perp \quad \eta_e \sim \text{Beta}^*(\mu_e, \mu_0 n_0) \quad \perp\!\!\!\perp \quad \eta_s \sim \text{Beta}^*(\mu_s, (1-\mu_0)n_0), \quad (4.28)$$

which is equivalent to the BREASE prior imposing a restriction on the choice of prior sample sizes n_e and n_s. Marginally, we also have

$$\theta_1 \sim \text{Beta}^*\left((1-\mu_e)\mu_0 + \mu_s(1-\mu_0), n_0\right). \quad (4.29)$$

4.3.5 Posterior sampling

Exact sampling

The posterior under (4.10) is given by the following mixture of independent betas[12]

$$\begin{aligned}\pi(\theta_0, \eta_e, \eta_s | \mathcal{D}) \propto \sum_{j=0}^{y_1} \sum_{k=0}^{N_1 - y_1} \binom{y_1}{j} \binom{N_1 - y_1}{k} &\times \theta_0^{y_0+j+k+\mu_0 n_0} (1-\theta_0)^{N-(y_0+j+k)+(1-\mu_0)n_0} \\ &\times \eta_e^{k+\mu_e n_e} (1-\eta_e)^{j+(1-\mu_e)n_e} \\ &\times \eta_s^{y_1-j+\mu_s n_s} (1-\eta_s)^{N_1-y_1-k+(1-\mu_s)n_s}. \end{aligned} \quad (4.30)$$

[12]Here $\text{Beta}(x; a, b)$ denotes the density of the $\text{Beta}(a, b)$ distribution evaluated at $x \in [0, 1]$.

Algorithm 1 BREASE posterior—exact sampling algorithm

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters $(\mu_0, \mu_e, \mu_s, n_0, n_e, n_s)$, and desired number of posterior samples T.

Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $P_1 \in \{0, \ldots, N_1 - y_1\}$ conditional on $\mathcal{D}$ with probability, according to (4.32),

$$\pi(P_1|\mathcal{D}) = \sum_{C_1=0}^{y_1} \pi(C_1, P_1|\mathcal{D}).$$

(ii) Sample $C_1 \in \{0, \ldots, y_1\}$ conditional on $(P_1, \mathcal{D})$ with probability, according to (4.32),

$$\pi(C_1|P_1, \mathcal{D}) \propto \pi(C_1, P_1|\mathcal{D}).$$

(iii) Sample $(\theta_0, \eta_e, \eta_s)$ conditional on $(C_1, P_1, \mathcal{D})$ from the distribution (4.33).

Output: Posterior samples $\{(\theta_0^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})\}_{t\in\{1,\ldots,T\}}$.

As with the prior, this posterior falls into the family of generalized Dirichlet distributions on the vector of potential outcomes probabilities $\mathbf{p}$. While some posterior quantities can be obtained analytically (see Section 4.3.7), working with the posterior density can often be cumbersome; thus, we now describe how to sample exactly from the posterior via simulation.

Theorem 2. *Let $(\theta_0, \eta_e, \eta_s)$ be random variables drawn according to Algorithm 1. Then $(\theta_0, \eta_e, \eta_s)$ are distributed according to the BREASE posterior (4.30).*

Before proceeding to the proof of Theorem 2, we motivate the argument with a sketch. We define the counterfactual counts

$$C_1 = \sum_{i=1}^{N} I(Z_i = 1, Y_i(1) = 1, Y_i(0) = 0), \qquad P_1 = \sum_{i=1}^{N} I(Z_i = 1, Y_i(1) = 0, Y_i(0) = 1),$$

which are unobserved quantities. Here, C_1 is the number of "causal" subjects in the treatment group, i.e., those who died under treatment but would have survived if untreated. Similarly,

P_1 is the number of "preventive" subjects in the treatment group, i.e., those who survived under treatment but would have died if untreated. The BREASE posterior can then be expressed as a mixture distribution:

$$\pi(\theta_0, \eta_e, \eta_s|\mathcal{D}) = \sum_{C_1=0}^{y_1} \sum_{P_1=0}^{N_1-y_1} \pi(\theta_0, \eta_e, \eta_s|C_1, P_1, \mathcal{D}) \times \pi(C_1, P_1|\mathcal{D}). \tag{4.31}$$

Hence, we can sample from the posterior by first drawing from the distribution of unobserved counts (C_1, P_1) conditional on the observed data $\mathcal{D}$. This distribution has probability mass function

$$\begin{aligned}\pi(C_1, P_1|\mathcal{D}) \propto & \binom{y_1}{C_1}\binom{N_1-y_1}{P_1}\mathrm{B}(P_1 + \mu_e n_e, y_1 - C_1 + (1-\mu_e)n_e) \\ & \times \mathrm{B}(y_0 + y_1 - C_1 + P_1 + \mu_0 n_0, N - (y_0 + y_1 - C_1 + P_1) + (1-\mu_0)n_0) \\ & \times \mathrm{B}(C_1 + \mu_s n_s, N_1 - y_1 - P_1 + (1-\mu_s)n_s).\end{aligned} \tag{4.32}$$

We then sample the parameters $(\theta_0, \eta_e, \eta_s)$, which have an independent beta distribution conditional on the augmented data $(C_1, P_1, \mathcal{D})$:

$$\begin{aligned}\pi(\theta_0, \eta_e, \eta_s|C_1, P_1, \mathcal{D}) = & \mathrm{Beta}(\eta_e; P_1 + \mu_e n_e, y_1 - C_1 + (1-\mu_e)n_e) \\ & \times \mathrm{Beta}(\theta_0; y_0 + y_1 - C_1 + P_1 + \mu_0 n_0, N - (y_0 + y_1 - C_1 + P_1) + (1-\mu_0)n_0) \\ & \times \mathrm{Beta}(\eta_s; C_1 + \mu_s n_s, N_1 - y_1 - P_1 + (1-\mu_s)n_s).\end{aligned} \tag{4.33}$$

Note that this derivation of the distribution (4.32) provides a counterfactual interpretation of the mixture weights that result from directly normalizing the kernels in (4.30).

Proof of Theorem 2. For treatment or control group $j \in \{1, 0\}$, respectively, define the re-

sponse type (doomed, immune, preventive, and causal) counts in that group as

$$\begin{aligned} D_j &= \sum_{i=1}^{N} I(Z_i = j, Y_i(0) = 1, Y_i(1) = 1), \\ I_j &= \sum_{i=1}^{N} I(Z_i = j, Y_i(0) = 0, Y_i(1) = 0), \\ P_j &= \sum_{i=1}^{N} I(Z_i = j, Y_i(0) = 1, Y_i(1) = 0), \\ C_j &= \sum_{i=1}^{N} I(Z_i = j, Y_i(0) = 0, Y_i(1) = 1). \end{aligned}$$

As noted above, the BREASE posterior can be expressed as the mixture distribution (4.31). We will derive each term in the sum. A straightforward calculation shows that

$$\begin{aligned} &(D_j, I_j, P_j, C_j)|(\theta_0, \eta_e, \eta_s, N_j) \sim \\ &\qquad \text{Multinomial}_{N_j}(\theta_0(1-\eta_e), (1-\theta_0)(1-\eta_s), \theta_0\eta_e, (1-\theta_0)\eta_s), \qquad j \in \{0, 1\}, \end{aligned}$$

and the two distributions are independent: $(D_0, I_0, P_0, C_0) \perp\!\!\!\perp (D_1, I_1, P_1, C_1)$. Since

$$y_1 = D_1 + C_1 \qquad \text{and} \qquad N_1 - y_1 = I_1 + P_1,$$

it follows that

$$\begin{aligned} C_1|(y_1, \theta_0, \eta_e, \eta_s) &\sim \text{Binomial}\left(y_1, \frac{(1-\theta_0)\eta_s}{\theta_1}\right), \\ P_1|(y_1, N_1, \theta_0, \eta_e, \eta_s) &\sim \text{Binomial}\left(N_1 - y_1, \frac{\theta_0\eta_e}{1-\theta_1}\right), \end{aligned}$$

independently. Consequently, we have

$$
\begin{aligned}
&\pi(\theta_0, \eta_e, \eta_s | C_1, P_1, \mathcal{D}) \\
&\quad \propto \pi(C_1, P_1, \mathcal{D} | \theta_0, \eta_e, \eta_s) \times \pi(\theta_0, \eta_e, \eta_s) \\
&\quad = \pi(C_1, P_1 | \mathcal{D}, \theta_0, \eta_e, \eta_s) \times \pi(\mathcal{D} | \theta_0, \eta_e, \eta_s) \times \pi(\theta_0, \eta_e, \eta_s) \\
&\quad = \pi(C_1 | y_1, \theta_0, \eta_e, \eta_s) \times \pi(P_1 | y_1, N_1, \theta_0, \eta_e, \eta_s) \\
&\qquad \times \pi(\mathcal{D} | \theta_0, \eta_e, \eta_s) \times \pi(\theta_0, \eta_e, \eta_s) \\
&\quad = \text{Binomial}\left(C_1; y_1, \frac{(1-\theta_0)\eta_s}{\theta_1}\right) \times \text{Binomial}\left(P_1; N_1 - y_1, \frac{\theta_0 \eta_e}{1-\theta_1}\right) \\
&\qquad \times \text{Binomial}(y_0; N_0, \theta_0) \times \text{Binomial}(y_1; N_1, \theta_1) \\
&\qquad \times \text{Beta}(\theta_0; \mu_0 n_0, (1-\mu_0) n_0) \times \text{Beta}(\eta_e; \mu_e n_e, (1-\mu_e) n_e) \times \text{Beta}(\eta_s; \mu_s n_s, (1-\mu_s) n_s) \\
&\quad \propto \theta_0^{y_0 + y_1 - C_1 + P_1 + \mu_0 n_0 - 1} (1-\theta_0)^{N - (y_0 + y_1 - C_1 + P_1) + (1-\mu_0) n_0 - 1} \\
&\qquad \times \eta_e^{P_1 + \mu_e n_e - 1} (1-\eta_e)^{y_1 - C_1 + (1-\mu_e) n_e - 1} \\
&\qquad \times \eta_s^{C_1 + \mu_s n_s - 1} (1-\eta_s)^{N_1 - y_1 - P_1 + (1-\mu_s) n_s - 1}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\pi(\theta_0, \eta_e, \eta_s | C_1, P_1, \mathcal{D}) \\
&\quad = \text{Beta}(\theta_0; y_0 + y_1 - C_1 + P_1 + \mu_0 n_0, N - (y_0 + y_1 - C_1 + P_1) + (1-\mu_0) n_0) \\
&\qquad \times \text{Beta}(\eta_e; P_1 + \mu_e n_e, y_1 - C_1 + (1-\mu_e) n_e) \\
&\qquad \times \text{Beta}(\eta_s; C_1 + \mu_s n_s, N_1 - y_1 - P_1 + (1-\mu_s) n_s). \qquad (4.34)
\end{aligned}
$$

Similarly, for the mixture weights we have

$$\begin{aligned}
\pi(C_1, P_1|\mathcal{D}) &= \int \pi(C_1, P_1, \theta_0, \eta_e, \eta_s|\mathcal{D}) d\theta_0 d\eta_e d\eta_s \\
&= \int \pi(C_1, P_1|\theta_0, \eta_e, \eta_s, \mathcal{D}) \pi(\theta_0, \eta_e, \eta_s|\mathcal{D}) d\theta_0 d\eta_e d\eta_s \\
&\propto \binom{y_1}{C_1} \binom{N_1 - y_1}{P_1} \mathrm{B}(P_1 + \mu_e n_e, y_1 - C_1 + (1-\mu_e) n_e) \\
&\quad \times \mathrm{B}(y_0 + y_1 - C_1 + P_1 + \mu_0 n_0, N - (y_0 + y_1 - C_1 + P_1) + (1-\mu_0) n_0) \\
&\quad \times \mathrm{B}(C_1 + \mu_s n_s, N_1 - y_1 - P_1 + (1-\mu_s) n_s). \qquad (4.35)
\end{aligned}$$

Hence, we can sample from the mixture distribution (4.31) using Algorithm 1.

□

Data augmentation (DA) algorithm

Although the exact sampler of Algorithm 1 is fast for most sample sizes in real-world settings, it may be useful to resort to MCMC when sampling from the discrete mixture weight distribution (4.32) is computationally demanding, e.g., when the size of the treatment arm N_1 and the number of cases y_1 is in the order of millions or higher. Thus we now derive a Gibbs sampler targeting the BREASE posterior (4.30) based on the data augmentation scheme introduced for Algorithm 1.

Algorithm 2 defines the Gibbs sampler. It consists of two steps: (i) first we sample the counterfactual counts C_1 and P_1 conditionally on θ_0 and θ_1; and, (ii) we sample θ_0, η_e, η_s conditionally on the augmented data. In numerical experiments, we find that the algorithm converges to the BREASE posterior rather quickly, often mixing within a few hundred iterations, and the sampling is also quite fast. The conditional distribution (4.36) of the unobserved counts $(C_1, P_1)|(\theta_0, \eta_e, \eta_s, \mathcal{D})$ was derived above in the proof of Theorem 2.

Algorithm 2 BREASE posterior—data augmentation algorithm

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters $(\mu_0, \mu_e, \mu_s, n_0, n_e, n_s)$, desired number of posterior samples T, number of burn-in iterations B, and BREASE parameter initialization $(\theta_0^{(0)}, \eta_e^{(0)}, \eta_s^{(0)}) \in (0,1)^3$.

Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $(C_1^{(t)}, P_1^{(t)})$ conditional on $(\theta_0^{(t-1)}, \eta_e^{(t-1)}, \eta_s^{(t-1)}, \mathcal{D})$ from the independent binomial distributions

$$C_1^{(t)} \sim \text{Binomial}\left(y_1, \frac{(1-\theta_0^{(t-1)})\eta_s^{(t-1)}}{\theta_1^{(t-1)}}\right),$$
$$P_1^{(t)} \sim \text{Binomial}\left(N_1 - y_1, \frac{\theta_0^{(t-1)}\eta_e^{(t-1)}}{1-\theta_1^{(t-1)}}\right), \tag{4.36}$$

where $\theta_1^{(t-1)} = \theta_0^{(t-1)}(1-\eta_e^{(t-1)}) + (1-\theta_0^{(t-1)})\eta_s^{(t-1)}$.

(ii) Sample $(\theta_0^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})$ conditional on $(C_1^{(t)}, P_1^{(t)}, \mathcal{D})$ from the independent beta distributions (4.33).

Output: Posterior samples after burn-in $\{(\theta_0^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})\}_{t \in \{B+1, \ldots, T\}}$.

Pathological sampling

To demonstrate the utility of our posterior sampling algorithms, we now turn to an example for which RJAGS (Plummer, 2023) and RStan (Stan Development Team, 2020), two popular MCMC software packages, fail to sample from the BREASE posterior. We use the data $y_0 = 20$, $N_0 = 1000$, $y_1 = 40$, $N_1 = 1000$, and the hyperparameters $\mu_0 = 0.5$, $n_0 = 2$, $\mu_e = 0.5$, $n_e = 2, \mu_s = 0.01$, $n_s = 1$. The prior distributions for θ_0 and η_e are vague independent Uniform$(0,1)$ distributions. On the other hand, the prior on the risk of side effects η_s is concentrated near 0 with mean $\mu_s = 0.01$. This prior encodes a quasi-monotonicity assumption on the treatment that is clearly in conflict with the data.

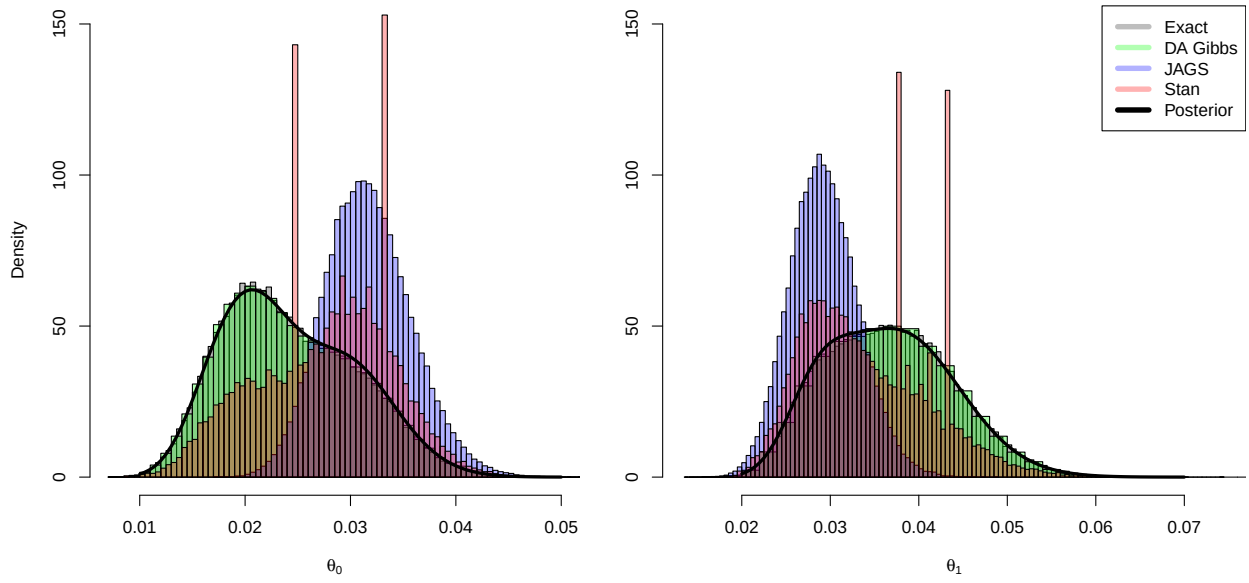


Figure 4.4: Pathological MCMC posterior sampling exhibited in posterior histograms of the baseline risk θ_0 (left) and treatment risk θ_1 (right). The marginal posterior of θ_1 (black curve) was approximated using numerical integration.

Prior-data conflict, which arises when the prior is concentrated on parameter values that are unlikely given the data, is a common culprit when diagnosing pathological MCMC sampling (Evans & Moshonov, 2006). This example is no exception. Figure 4.4 shows histograms of 100,000 posterior samples of θ_0 and θ_1 drawn using Algorithm 1 (grey), Algorithm 2 (green), JAGS (blue), and Stan (red). The marginal posterior density is plotted in black for reference. The posterior of θ_0 is a mixture of beta distributions and its multimodality is exhibited in the left panel of Figure 4.4. While Algorithms 1 and 2 produce posterior samples that fully capture the distribution, JAGS and Stan fail to adequately explore the left-hand mode. Although Stan manages to deviate from the right-hand mode as compared to JAGS, its chains get stuck at $\theta_0 \approx 0.024$ and $\theta_0 \approx 0.033$ when the sampler rejects numerous proposal

draws. The story is much the same for θ_1.

Perhaps troublingly, JAGS sounds no alarm and its diagnostics provide no indication that the MCMC hasn't converged. Stan, on the other hand, produces warning messages that the sampler is struggling. It reports a substantial proportion of divergent transitions, indicating that the chain has reached a portion of state space with extreme posterior curvature, which poses a challenge to gradient evaluation. Based on further investigation, the problem may be due to numerical issues in dealing with the Beta$^*(0.01, 1)$ prior on η_s, which is highly concentrated and diverges at the boundary $\eta_s = 0$. (Indeed, in our experiments, Stan struggles to sample from the Beta$^*(0.01, 1)$ distribution even in the absence of data and the other BREASE parameters.) There are a number of potential ways to ameliorate this issue. One is to express the posterior in the mixture form (4.30). In this representation we obtain a beta distribution more amenable to sampling by updating the prior on η_s with the counterfactual counts C_1, P_1 . When N_1 and y_1 are large, however, the double sum in (4.30) can become prohibitively expensive to evaluate, which slows MCMC sampling significantly. Alternatively, we may circumvent dealing with η_e and η_s entirely by marginalizing one of them out of the prior and carrying out a change of variables to evaluate the conditional prior $\pi(\theta_1|\theta_0)$ using numerical integration as in (4.15). In either case, however, we find that the resulting sampler remains less efficient and more challenging to implement (even with existing software) than the exact sampler of Algorithm 1 and the data-augmented Gibbs sampler of Algorithm 2. Nevertheless, this example demonstrates that it is useful to have bespoke algorithms that we know will perform well, even in adversarial settings. In future extensions of the model—e.g., if we account for covariates, expressing the BREASE parameters as functions of them on the logit scale—specialized MCMC algorithms tailored to binomial likelihoods, such as the Pólya-Gamma augmented Gibbs sampler of Polson et al. (2013), may provide better performance than off-the-shelf MCMC methods. We leave exploration of this for future work.

Monotonicity. Posterior sampling under monotonicity constraints, such as setting $\eta_s = 0$ or $\eta_e = 0$, can be obtained with similar procedures, and we thus defer their discussion to the appendix. See Theorems 4-5 of Appendix A.2.

4.3.6 Marginal likelihoods and Bayes factors

From a Bayesian perspective, hypothesis testing is essentially a model comparison exercise (Dickey & Lientz, 1970; Jeffreys, 1961; Kass & Raftery, 1995). Consider two competing hypothesis, H_0 and H_1. For each hypothesis H_k, $k \in \{0, 1\}$, the Bayesian approach requires postulating a fully specified model M_k, with likelihood $L_k(\mathcal{D}|\theta)$ and prior $\pi_k(\theta)$, respecting the constraints of the hypothesis the model is intended to represent. Evidence in favor of H_1 relative to H_0 is then quantified using the Bayes factor BF_{10}, given by the ratio of the marginal likelihoods of the observed data under each model, $\mathrm{BF}_{10} = L_1(\mathcal{D})/L_0(\mathcal{D})$, where $L_k(\mathcal{D}) = \int L_k(\mathcal{D}|\theta)\pi_k(\theta)d\theta$. Given prior model probabilities $\mathbb{P}(M_0)$, $\mathbb{P}(M_1)$, the posterior odds of M_1 and M_0 are then $\mathbb{P}(M_1|\mathcal{D})/\mathbb{P}(M_0|\mathcal{D}) = \mathrm{BF}_{10} \times \mathbb{P}(M_1)/\mathbb{P}(M_0)$. In this section we show how to formulate such models instantiating a number of relevant statistical hypotheses with the BREASE approach, and provide analytical formulae for the marginal likelihoods. For all models considered here the likelihood is the same, so we focus the discussion on the formulation of the prior.

Let us first consider testing the null hypothesis $H_0 : \theta_1 = \theta_0$ against the alternative hypothesis $H_1 : \theta_1 \neq \theta_0$. For H_1, we propose using the unconstrained model M_1, with the BREASE prior in (4.10) and equation (4.8),

$$M_1 : (\theta_0, \eta_e, \eta_s) \sim \mathrm{BREASE}(\mu; n), \qquad \theta_1 = (1 - \eta_e)\theta_0 + \eta_s(1 - \theta_0). \tag{4.37}$$

As for the null hypothesis $H_0 : \theta_1 = \theta_0$, we instantiate it with the null model,

$$M_0 : \theta_0 \sim \mathrm{Beta}^*(\mu_0, n_0), \qquad \theta_1 = \theta_0. \tag{4.38}$$

One benefit of M_0 is that its prior is logically consistent with the marginal distribution of θ_0 under M_1, both implying $\theta_0 \sim \text{Beta}^*(\mu_0, n_0)$ *a priori*. Note that the prior (4.38) emerges naturally from M_1 in at least two ways: (i) when postulating that the treatment does not work at all, by setting $\eta_s = \eta_s = 0$; or, (ii) by noting that, if the treatment has no effect on average (i.e, the efficacy of the treatment precisely offsets its side effects), one can side-step thinking about η_s and η_e altogether. In both cases, we borrow the prior of θ_0 from M_1, and simply set θ_1 equal to θ_0. We discuss alternative prior formulations for H_0 in Appendix A.3.1.

Other relevant hypothesis one may wish to test are that the treatment is beneficial $H_- : \theta_1 < \theta_0$ or that the treatment is harmful $H_+ : \theta_1 > \theta_0$, on average. A straightforward approach to specify models for such hypotheses is to note that M_1 already induces positive probabilities to the events postulated in H_- and H_+. Thus, we can borrow this knowledge, already elicited when forming M_1, to define the priors π_- and π_+,

$$\pi_-(\theta_0, \eta_e, \eta_s) := \pi_1(\theta_0, \eta_e, \eta_s | \theta_1 < \theta_0), \quad \pi_+(\theta_0, \eta_e, \eta_s) := \pi_1(\theta_0, \eta_e, \eta_s | \theta_1 > \theta_0). \tag{4.39}$$

The priors π_- and π_+ result in the models M_- and M_+, for H_- and H_+ respectively. Similarly to M_0, one benefit of these models is that the induced priors on $(\theta_0, \eta_e, \eta_s)$ are logically consistent with the beliefs expressed in M_1, under the constraints H_- and H_+. Note that the same strategy employed here can be used for interval hypotheses of the type $H_0^\delta : |\theta_1 - \theta_0| \leq \delta$, with $\delta > 0$ (or, more generally, for any event with nonzero probability under M_1). Alternative models for H_- and H_+, leveraging instead monotonicity constraints, such as $\eta_s = 0$ or $\eta_e = 0$, are discussed in Appendix A.3.2.

In all cases above, the marginal likelihood can be obtained using analytical formulae and simple Monte Carlo approximation, thereby facilitating the computation of Bayes factors.

Theorem 3. *The marginal likelihood of the data under M_0 is given by a beta-binomial*

distribution. Under M_1, it is given by a weighted sum of beta functions:[13]

$$L_1(\mathcal{D}) = \binom{N_0}{y_0}\binom{N_1}{y_1}\sum_{j=0}^{y_1}\sum_{k=0}^{N_1-y_1}\binom{y_1}{j}\binom{N_1-y_1}{k} \times \frac{B(k+\mu_e n_e, j+(1-\mu_e)n_e)}{B(\mu_e n_e,(1-\mu_e)n_e)} \times \frac{B(y_0+j+k+\mu_0 n_0, N-(y_0+j+k)+(1-\mu_0)n_0)}{B(\mu_0 n_0,(1-\mu_0)n_0)} \times \frac{B(y_1-j+\mu_s n_s, N_1-y_1-k+(1-\mu_s)n_s)}{B(\mu_s n_s,(1-\mu_s)n_s)}. \tag{4.40}$$

Under M_- and M_+, it can be obtained from $L_1(\mathcal{D})$ as follows,

$$L_-(\mathcal{D}) = L_1(\mathcal{D}) \times \frac{\pi_1(\theta_1 < \theta_0|\mathcal{D})}{\pi_1(\theta_1 < \theta_0)}, \qquad L_+(\mathcal{D}) = L_1(\mathcal{D}) \times \frac{\pi_1(\theta_1 > \theta_0|\mathcal{D})}{\pi_1(\theta_1 > \theta_0)}. \tag{4.41}$$

Proof. The result for M_0 is well-known. $L_1(\mathcal{D})$ in (4.40) follows directly from integration of (4.9) under the prior (4.10). $L_-(\mathcal{D})$ and $L_+(\mathcal{D})$ in (4.41) follow from Bayes' rule. □

Remark 1. The prior and posterior probabilities $\pi_1(\theta_1 < \theta_0)$ and $\pi_1(\theta_1 < \theta_0|\mathcal{D})$ can be approximated using Monte Carlo integration with exact samples, as per Section 4.3.5.

Remark 2. As per (4.41), if one postulates prior model probabilities $\mathbb{P}(M_-|M_1) = \pi_1(\theta_1 < \theta_0)$ and $\mathbb{P}(M_+|M_1) = \pi_1(\theta_1 > \theta_0)$, the Bayes factor testing $H_0 : \theta_1 = \theta_0$ against $H_1 : \theta_1 \neq \theta_0$ (using M_1) conveniently decomposes into the weighted average of the Bayes factors testing H_0 against H_- (using M_-) and H_0 against H_+ (using M_+)—though, of course, users can postulate prior probabilities for the models M_- and M_+ as they wish.

As noted by Campbell and Gustafson (2022), if one reports a Bayes factor comparing models, it is advisable to also report posterior estimates accounting for model uncertainty, i.e., using the implied mixture prior given by the weighted combination of the priors of all models being compared, $\pi(\theta) = \sum_k \mathbb{P}(M_k)\pi_k(\theta)$. In this case, samples from the mixture posterior can be readily obtained by sampling from the posterior of each model (as detailed in Section 4.3.5) proportionally to each model's posterior probability, $\pi(\theta|\mathcal{D}) = \sum_k \mathbb{P}(M_k|\mathcal{D})\pi_k(\theta|\mathcal{D})$.

[13]Here $\mathrm{B}(a, b)$ denotes the beta function evaluated at (a, b).

4.3.7 Posterior quantities of interest

In addition to marginal likelihoods, we can derive analytical expressions for certain relevant functionals of the BREASE posterior distribution $\pi(\theta_0, \eta_e, \eta_s|\mathcal{D})$. While posterior quantities can generally be easily estimated using simple Monte Carlo approximation with samples obtained from Algorithm 1, analytical formulae may be of value, e.g., for conducting prior sensitivity analysis of treatment effect estimands without needing to sample the posterior for every choice of the hyperparameters (μ, n).

The risk difference $\theta_1 - \theta_0$ and risk ratio θ_1/θ_0 are of particular interest in practice, with expectations of their posterior distributions often reported. We first note that, since the posterior $\pi(\theta_0, \eta_e, \eta_s|\mathcal{D})$ is a mixture of independent beta distributions, conditional and marginal expectations and percentiles can be easily computed by first calculating expectations or percentiles of the beta summands and averaging these quantities across the mixture weights. For example, using the mixture representation (4.31) of the posterior, we have

$$\begin{aligned}\mathbb{E}[\theta_0|\mathcal{D}] &= \int \theta_0 \cdot \pi(\theta_0, \eta_e, \eta_s|\mathcal{D}) d\theta_0 d\eta_e d\eta_s \\ &= \sum_{y_1(0)=0}^{y_1} \sum_{x_1(1)=0}^{N_1-y_1} \pi(y_1(0), x_1(1)|\mathcal{D}) \int \theta_0 \cdot \pi(\theta_0, \eta_e, \eta_s|y_1(0), x_1(1), \mathcal{D}) d\theta_0 d\eta_e d\eta_s.\end{aligned}$$

Applying equations (4.32) and (4.33) then yields an expression for $\mathbb{E}[\theta_0|\mathcal{D}]$ in terms of the data $\mathcal{D}$ and hyperparameters (μ, n), which we omit for brevity. In a similar fashion, by exploiting the mixture-of-betas representation of the posterior, we can easily calculate posterior expectations of polynomials $\sum_{(\alpha_0,\alpha_e,\alpha_s)} a_{(\alpha_0,\alpha_e,\alpha_s)} \theta_0^{\alpha_0} \eta_e^{\alpha_e} \eta_s^{\alpha_s}$, including those with negative exponents, assuming $\mathcal{D}$ and (μ, n) are such that the integrals converge.

In particular, assuming treatment is not harmful ($\eta_s = 0$), the efficacy can be written in terms of the risk ratio as $\eta_e = 1 - \theta_1/\theta_0$. The formulae derived in Appendix A.2.1 can then be applied to calculate $\mathbb{E}[\theta_1/\theta_0|\mathcal{D}] = 1 - \mathbb{E}[\eta_e|\mathcal{D}]$ using the posterior $\pi(\theta_0, \eta_e|\mathcal{D})$ under the

monotonicity assumption. More generally, we have

$$\begin{aligned}\mathbb{E}[\theta_1/\theta_0|\mathcal{D}] &= \mathbb{E}\left[\frac{\theta_0(1-\eta_e-\eta_s)+\eta_s}{\theta_0}\middle|\mathcal{D}\right] \\ &= 1-\mathbb{E}[\eta_e|\mathcal{D}]-\mathbb{E}[\eta_s|\mathcal{D}]+\mathbb{E}[\theta_0^{-1}\eta_s|\mathcal{D}].\end{aligned}$$

Similarly, the expected posterior risk difference can be obtained as

$$\mathbb{E}[\theta_1-\theta_0|\mathcal{D}] = \mathbb{E}[\eta_s|\mathcal{D}]-\mathbb{E}[\theta_0\eta_e|\mathcal{D}]-\mathbb{E}[\theta_0\eta_s|\mathcal{D}].$$

In Section 4.4 we demonstrate how to conduct sensitivity analysis with the BREASE prior for Bayes factors using the marginal likelihoods derived in Section 4.3.6. The discussion therein applies just as well to treatment effects and other posterior quantities.

4.4 Empirical Examples

We now demonstrate the utility of our approach in three empirical examples. We show how the BREASE framework can be used to facilitate Bayesian estimation, hypothesis testing, and sensitivity analysis of the results of binary experiments. Concretely, the examples illustrate how our proposal can: (i) help analysts distinguish robust from fragile findings; (ii) clarify what one needs to believe in order to claim that a treatment is effective; and (iii) reconcile disparate results obtained from different methods.

4.4.1 Implementation

Following Dablander et al. (2022), we calculate the Bayes factor BF_{10} for the IB approach using the Savage-Dickey density ratio method applied to the difference of proportions $\eta = \theta_0 - \theta_1$ (Wagenmakers et al., 2010). A formula for the prior density of η at the null $H_0 : \eta = 0$ can be found in Appendix A of (Dablander et al., 2022). The Bayes factor using the $\text{IB}((a,a),(a,a))$ prior under H_1 as described in Section 4.2.2 is then

$$\text{BF}_{10} = \frac{\text{B}(2a-1,2a-1)\text{B}(a+y_0,a+N_0-y_0)\text{B}(a+y_1,a+N_1-y_1)}{\text{B}(2a+y_0+y_1-1,2a+N_0-y_0+N_1-y_1-1)\text{B}(a,a)^2}.$$

Posterior estimates and credible intervals under H_1 are calculated using exact samples from the independent beta posterior.

Bayes factors BF_{10} for the LT approach are calculating using the **abtest** package in R (Gronau et al., 2021). The package uses a Laplace approximation to calculate BF_{10}, which is shown to have good performance. The LT prior under H_1 is as described in Section 4.2.2. Under $H_0 : \psi = 0$, the prior is $\beta \sim \text{Normal}(\mu_\beta, \sigma_\beta)$ with default values $(\mu_\beta, \sigma_\beta) = (0, 1)$. Posterior estimates and credible intervals under H_1 are calculated using posterior samples output by **abtest**. As **abtest** only reports marginal likelihoods up to a multiplicative constant, we used RJAGS (Plummer, 2023) to generate MCMC samples from the LT posterior and THAMES (Metodiev et al., 2024) to estimate the LT marginal likelihood for Figure 4.7b using the samples.

4.4.2 The effect of aspirin on myocardial infarction

We revisit the aspirin component of the Physicians' Health Study, a large-scale randomized, placebo-controlled trial designed, in part, to investigate whether low-dose aspirin decreases the risk of cardiovascular mortality (Steering Committee of the Physicians' Health Study Research Group, 1989). During the study, $y_0 = 26$ out of $N_0 = 11{,}034$ subjects in the placebo group experienced fatal myocardial infarction compared to $y_1 = 10$ out of $N_1 = 11{,}037$ prescribed aspirin. Using maximum likelihood estimation, the estimated risk ratio θ_1/θ_0 is 0.38, with 95% confidence interval (based on inverting Fisher's exact test) $\text{CI}(95\%) = [0.17, 0.82]$. Consequently, we reject the null hypothesis of zero effect, $H_0 : \theta_1 = \theta_0$, with p-value 0.008. Results based on asymptotic Wald and Pearson tests are nearly identical. Hence, a frequentist would confidently conclude that low-dose aspirin significantly reduces cardiovascular mortality.

Traditional Bayesian estimation under the alternative hypothesis (i.e, with a prior that gives zero probability to the null hypothesis of zero effect) yields qualitatively similar, though

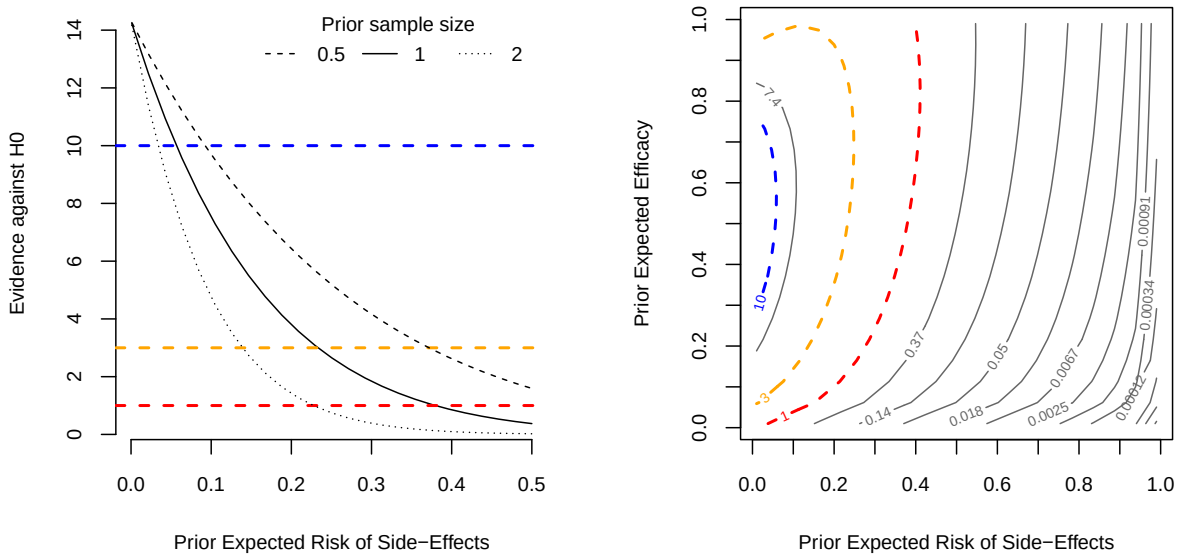
more conservative, answers. Using our default prior, BREASE$(1/2, .3, .3; 2, 1, 1)$, the posterior median of the risk ratio is 0.44 with a wider 95% credible interval of CrI(95%) = $[0.2, 0.96]$. The results for the LT and IB approach are similar.[14]

Traditional estimation, however, does not give the null hypothesis of zero effect a fighting chance, as it is assumed to be false *a priori*. One may thus be interested in performing a Bayesian hypothesis test assigning nonzero prior probability to H_0.[15] Perhaps surprisingly, a test based on the IB approach yields a Bayes factor $\text{BF}_{01} = 20.27$, suggesting that the data provide strong evidence *in favor* of H_0. On the other hand, the Bayes factor under the LT approach is $\text{BF}_{10} = 5.24$, which suggests moderate evidence in favor of $H_1 : \theta_1 \neq \theta_0$. Finally, the default BREASE prior results in $\text{BF}_{10} = 1.2$ providing essentially little evidence in favor of one hypothesis or the other. How can we make sense of these disparate results? As is well known, Bayes factors are sensitive to the prior distribution (Kass & Raftery, 1995). It is important, then, that prior assumptions are encoded in a way that practitioners can understand, both to examine the reasonableness of the prior, as well as to explore how robust inferences are to sensible perturbations of the prior (Gunel, 1984; Kass & Raftery, 1995; Leamer, 1978).

One benefit of the BREASE approach is that it allows one to clearly encode prior assumptions in terms of the expected efficacy and side effects of aspirin, and to examine how sensitive the BF is to those assumptions. For example, aspirin is an over-the-counter medicine, with ample usage, and it would thus be unreasonable to expect that aspirin would *cause* myocardial infarction in a large fraction of otherwise healthy patients. Figure 4.5a inspects how the Bayes factor is affected as we vary the prior expectation of side effects, ranging from 0.01% to 50%, while still keeping relatively vague priors on the baseline risk and efficacy. The

[14] LT(0,0;1,1): median = 0.48 and CrI(95%) = $[0.25, 0.87]$. IB(1,1;1,1): median = 0.4 and CrI(95%) = $[0.18, 0.79]$.

[15] Here we focus on the exact null, but we note that researchers can also specify an interval null hypothesis, such as $|\theta_1 - \theta_0| < \delta$, as per discussion of Section 4.3.6.



(a) Sensitivity plot varying μ_s and n_s.

(b) Sensitivity contours varying μ_s and μ_e.

Figure 4.5: Sensitivity analysis of BF_{10} for the aspirin trial.

dashed red, orange, and blue lines denote (slightly modified) Jeffreys' thresholds for weak ($1 \leq \text{BF}_{10} \leq 3$), moderate ($3 \leq \text{BF}_{10} \leq 10$), and strong ($\text{BF}_{10} \geq 10$) evidence against H_0, respectively (Jeffreys, 1961; Kass & Raftery, 1995). Indeed, as the plot shows, the results are extremely sensitive to the choice of μ_s. Setting the expected value of side effects to 1% results in $\text{BF}_{10} = 13.45$, yielding strong evidence in favor of H_1, while setting it to 50% results in $\text{BF}_{01} = 2.66$, yielding weak evidence in favor of H_0. Translating these to posterior probabilities, we have the wide range of 27% to 93% probability of the existence of an effect (assuming equal prior odds for H_0 and H_1).

One may also want to conduct a sensitivity analysis with respect to both hyperparameters simultaneously for the $\text{BREASE}(1/2, \mu_e, \mu_s; 2, 1, 1)$ prior. Figure 4.5b shows the contour lines of BF_{10} as a function of $(\mu_e, \mu_s) \in (0, 1)^2$ over their full range of possible values, while keeping

$n_e = n_s = 1$ fixed. In general, the results seem more sensitive to plausible variations of the expected risk of side effects μ_s than to plausible variations of the expected efficacy μ_e of aspirin. Overall, only when (i) side effects are expected to be small ($< 1\%$), and (ii) the efficacy is expected to be relatively large (between 30% and 70%), does the Bayes factor provide strong evidence against the null of no effect. For all other combinations of prior hyperparameters, the evidence is either moderate, weak, or favors the null. In this light, the results of the trial are ambiguous, and the conclusion that aspirin prevents heart attack strongly depends on the prior. Note that this need not always be the case, as we show below in a reanalysis of the Pfizer-BioNTech COVID-19 vaccine trial.

Finally, to assess the influence of prior correlation on the results and how they vary across methods, we compare estimates from the IB, LT, and BREASE priors with nearly matching marginals on the aspirin data. As noted above, the default IB prior, which places independent uniform distributions on θ_0 and θ_1 yields an estimated risk ratio RR = 0.4 [0.18, 0.79] under H_1 and Bayes factor $\text{BF}_{01} = 20.27$. The corresponding estimates under the default BREASE prior are RR = 0.44 [0.2, 0.96] and $\text{BF}_{10} = 1.2$. Under the LT prior calibrated to have near-uniform marginals and prior correlation 0.4 (described in Section 4.3.5), we have RR = 0.44 [0.21, 0.82] and $\text{BF}_{10} = 4.9$. Finally, under the LT prior calibrated to have near-uniform marginals and concentrate mass on the diagonal $\theta_0 = \theta_1$, we have RR = 0.93 [0.77, 1.12] and $\text{BF}_{10} = 1.3$. Hence, keeping the marginals fixed (or nearly fixed), we see variation in the Bayes factor across two orders of magnitude by varying the correlation structure between θ_0 and θ_1 across the IB, BREASE, and LT priors considered here. Note that this variation across methods is comparable to that within the default BREASE Bayes factor alone as we vary $\mu \in (0, 0.5)$, yielding prior correlation $1 - 2\mu \in (0, 1)$, as Figure 4.5b demonstrates. Hence, it appears that the correlation structure is a large source of variation in the Bayes factor (both within and across methods). We also observe that the LT prior that matches the default BREASE correlation yields a qualitatively similar posterior risk ratio but a Bayes factor

that now indicates substantial evidence in favor of H_1 (Kass & Raftery, 1995); on the other hand, the LT prior that concentrates mass on the diagonal yields a Bayes factor indicating equivocal evidence, similar to BREASE, but also a posterior risk ratio shrunk toward 1, which disagrees with the other methods. Hence, although we are able to match the Bayes factors of the LT prior to that of the default BREASE, the requisite LT hyperparameters could be considered extreme in some situations. In this example, the BREASE prior can be a useful complement to the LT by showing that it is indeed possible to get a wide range of Bayes Factors without using a prior that overwhelms the data, while still matching the marginal uniform priors of the IB and inducing positive prior correlation across study arms.

4.4.3 The Pfizer-BioNTech COVID-19 vaccine trial

We now reexamine the results of the Pfizer-BioNTech mRNA COVID-19 vaccine study (Polack et al., 2020). The experiment was a global multi-phase randomized placebo-controlled trial designed, in part, to evaluate the efficacy of the BNT162b2 vaccine candidate in preventing COVID-19. Vaccine development and evaluation were carried out in rapid response to the emerging SARS-CoV-2 pandemic. The results of the trial were definitive and precipitated the U.S. Food and Drug Administration's emergency use authorization for widespread dissemination of the vaccine (U.S. Food and Drug Administration, 2020).

During the study, $y_1 = 9$ out of $N_1 = 19{,}965$ subjects contracted COVID-19 subsequent to the second dose of the vaccine, while there were $y_0 = 169$ cases out of $N_0 = 20{,}172$ subjects receiving placebo injections. In their paper, Polack et al. adopted a Bayesian approach, focusing particularly on evaluating the vaccine's efficacy (defined in the study as the estimand $1 - \theta_1/\theta_0$). The efficacy of the vaccine was estimated at 0.95, with credible interval $\text{CrI}(95\%) = [0.90, 0.97]$. Frequentist estimates are similar, with a point estimate of 0.95, confidence interval $\text{CI}(95\%) = [0.90, 0.97]$, and a p-value for testing the null hypothesis of zero effect of the order 6×10^{-33}.

Polack et al. (2020) estimate $1 - \theta_1/\theta_0$ as the efficacy of the vaccine, but, as per Section 4.3.1, this only has the counterfactual interpretation of efficacy (i.e., $\eta_e = 1 - \theta_1/\theta_0$) under the assumption of monotonicity. Using the BREASE approach we can easily encode the monotonicity assumption by setting $\eta_s = 0$ and then proceed with estimation. The default BREASE prior, with the monotonicity constraint, results in posterior median and 95% credible interval for $\eta_e = 1 - \theta_1/\theta_0$ that are essentially the same as the previous results, namely, 0.94 and CrI(95%) = $[0.90, 0.97]$. In the absence of the monotonicity assumption, we have that $1 - \theta_1/\theta_0$ is in fact a lower bound on η_e. Again using the default BREASE prior, results are virtually unchanged, with posterior median and 95% credible interval for $1 - \theta_1/\theta_0$ of 0.94 and CrI(95%) = $[0.90, 0.97]$.[16] Conclusions using the IB and LT priors are practically equivalent.[17]

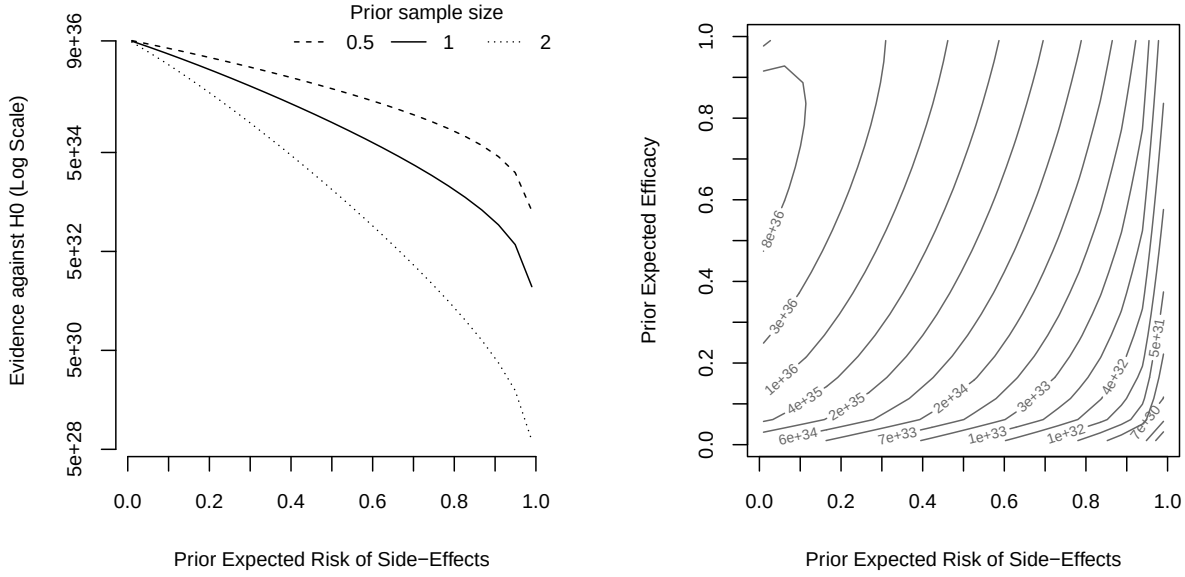
Turning to hypothesis testing, differently from the aspirin study, here all approaches point to the same direction, with overwhelming evidence against H_0. The Bayes factors against the null hypothesis of zero effect are 9×10^{33}, 5×10^{34} and 4×10^{35} for the IB, LT and BREASE default priors, respectively. Further, sensitivity analyses reveal the Bayes factor is in fact robust to variations in the hyperparameters across the whole range of prior expected efficacy and side effects of the vaccine, i.e., $(\mu_e, \mu_s) \in (0, 1)^2$. Figure 4.6 replicates the same sensitivity plots of the aspirin study for the COVID-19 trial. Notice that, in all scenarios, the posterior probability of the null hypothesis is essentially zero even if we posit equal prior odds for H_0 and H_1. Therefore, in this case, credible intervals constructed under H_1, neglecting H_0, are identical to credible intervals constructed using the mixture prior assigning a point mass of 0.5 to H_0. The trial provides unequivocal evidence that the vaccine is highly efficacious.

Finally, to assess the influence of prior correlation on the results and how they vary

---

[16] Corresponding values for η_e are 0.96 and CrI(95%) = $[0.90, 0.99]$. In this case, however, since η_e is not identified, the posterior of η_e is sensitive to the prior, and it remains spread in the partially identified region of η_e regardless of sample size.

[17] LT(0,0;1,1): med = 0.91, CrI(95%) = $[0.86, 0.95]$. IB(1,1;1,1): med = 0.94, CrI(95%) = $[0.90, 0.97]$.



(a) Sensitivity plot varying μ_s and n_s.

(b) Sensitivity contours varying μ_s and μ_e.

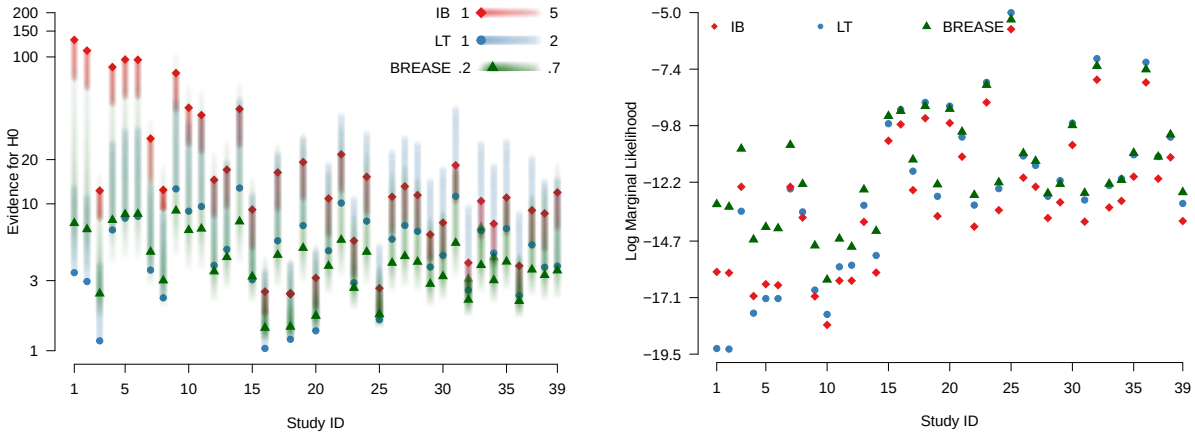
Figure 4.6: Sensitivity analysis of BF_{10} for the COVID-19 vaccine trial.

across methods, as in the previous example, we compare estimates from the IB, LT, and BREASE priors with nearly matching marginals on the COVID data. As noted above, the default IB prior yields an estimated vaccine efficacy VE = 0.94 [0.90, 0.97] under H_1 and Bayes factor $\text{BF}_{10} = 9 \times 10^{33}$. The corresponding estimates under the default BREASE prior are VE = 0.94 [0.90, 0.97] and $\text{BF}_{10} = 4 \times 10^{35}$. Under the LT prior calibrated to have near-uniform marginals and prior correlation 0.4, we have VE = 0.93 [0.88, 0.96] and $\text{BF}_{10} = 2 \times 10^{35}$. Finally, under the LT prior calibrated to have near-uniform marginals and concentrate mass on the diagonal $\theta_0 = \theta_1$, we have VE = 0.42 [0.32, 0.51] and $\text{BF}_{10} = 3 \times 10^{9}$. For all priors, the Bayes factors indicate decisive evidence in favor of H_1. Notably however, the LT prior that concentrates mass on the diagonal seems to exhibit large shrinkage when compared to the IB and BREASE estimates.

4.4.4 Null results in the New England Journal of Medicine

Dablander et al. (2022) conducted a Bayesian reanalysis of 39 binary experiments reporting null results (claiming absence or nonsignificance of an effect of treatment) in the *New England Journal of Medicine* (NEJM). They were particularly concerned with distinguishing between *absence of evidence* and *evidence of absence* of an effect when outcomes in the treatment and control groups are similar. Finding that Bayes factors calculated using the IB approach often strongly favored the null hypothesis (leaning heavily toward *evidence of absence*) whereas LT Bayes factors were generally equivocal, Dablander et al. concluded that the LT approach should be preferred for Bayesian tests for an equality of proportions. In our final empirical example, we expand their reanalysis to include the BREASE approach, and we show how it can easily address the concerns of Dablander et al. while also providing a better fit to the data in most cases.

Figure 4.7a contrasts the Bayes factors in favor of the null hypothesis using: (i) the IB$(a, a; a, a)$ prior varying $a \in [1, 5]$ (red diamonds); (ii) the LT$(0, 0; 1, \sigma_\psi)$ prior varying $\sigma_\psi \in [1, 2]$ (blue circles); and, the BREASE$(1/2, \mu, \mu; 2, 1, 1)$ prior varying $\mu \in [.2, .7]$ (green triangles). The solid color stands for the proposed default values of each method, namely $a = 1$ for the IB, $\sigma_\psi = 1$ for the LT and $\mu = .3$ for the BREASE. Note that the Bayes factors of the BREASE and LT default priors (solid triangle and circles) are similar across studies. Moreover, Dablander et al. (2022) noted that, in many examples, the Bayes factors of the IB and LT approaches could not be easily reconciled, even when reasonably varying their hyperparameters. The BREASE approach shows that this behavior is a mere artifact of those parameterizations. Indeed, for all studies, the BREASE prior easily interpolates between the two regimes, thus solving the apparent contradiction between the results of the LT and IB approaches, by transparently revealing how sensitive inferences are to the prior expected efficacy and side effects of the treatment μ. Finally, Figure 4.7b compares the predictive performance of the default IB, LT, and BREASE priors via the log marginal likelihood. The



(a) Bayes factors.

(b) Log marginal likelihoods.

Figure 4.7: Comparisons of log marginal likelihoods and Bayes factors across 39 NEJM studies, for the IB, LT and BREASE priors.

BREASE prior exhibits superior performance in *every study* when compared to the IB prior, and in more than 74% of the studies when compared to the LT prior. Thus, in this setting, our proposed default prior seems to provide not only a more sensible parameterization, but also a better fit to the data.

4.5 Discussion

We have introduced the BREASE framework for the Bayesian analysis of randomized controlled trials with a binary treatment and outcome. Framing the problem in the language of potential outcomes, we reparameterized the likelihood in terms of clinically meaningful quantities—the baseline risk, efficacy, and risk of adverse side effects of the treatment—and proposed a simple, yet flexible jointly independent beta prior distribution on these parame-

ters. We provided algorithms for exact posterior sampling, as well as analytical formulae for marginal likelihoods, Bayes factors, and other quantities. Finally, we showed with empirical examples how our proposal facilitates estimation, hypothesis testing, elicitation of prior knowledge and sensitivity analysis of treatment effects in binary experiments.

Chapter 5

DISCUSSION

The COVID-19 pandemic was a massive upheaval, the social impact of which we will continue to feel for decades to come. This dissertation documents the attempts of a few statisticians to make contributions to our understanding and response to COVID-19 and public health concerns more broadly.

During the prolonged duration of the pandemic, there has been a number of epidemiological and public health developments complicating estimation of SARS-CoV-2 incidence over time. These complicating factors include differential virulence of viral variants, the widespread administration of vaccines, improvements in medical treatment of COVID-19, and temporal variation in hospital capacity and case and death ascertainment, among others. In particular, key epidemiological parameters underlying models of SARS-CoV-2 transmission, such as the infection fatality rate (IFR), the basic reproduction number R_0, and the durations of the incubation, infectious, and immune periods—parameters often held fixed in modeling studies—have exhibited complex temporal dynamics. These considerations, combined with the inherent limitations of the most widely reported COVID-19 data—counts of deaths and positive cases—obscure the true number of SARS-CoV-2 infections incident over time. As we have demonstrated in this dissertation, knowledge of SARS-CoV-2 incidence is essential for accurately tracking viral spread, informing public policy, assessing the effectiveness of public health measures, and determining the burden of disease in a population, among other reasons. Going forward, it would be valuable, then, to incorporate more up-to-date data and extend the methods of Irons and Raftery (2021) in order to estimate COVID-19 prevalence over a longer time period accounting for these complexities. Measurements of

SARS-CoV-2 concentrations shed in wastewater are an important new source of information relevant to this task.

There are a number of interesting extensions of the BREASE framework enabling principled, interpretable, and transparent inference of treatment effects in experimental and observational studies with a dichotomous outcome, which are ubiquitous in the health and social sciences. One possibility is to extend the method to pool evidence across multiple trials. The problem of aggregating evidence is important in its own right, and data from multiple sites may also allow to point identify, or at least narrow the bounds on the fraction of people who benefit or are harmed by the intervention. As a Bayesian model specified explicitly in terms of the effects of treatment and the risk in the control group, BREASE is particularly well suited to the pooled analysis of multiple studies and the generalization of experimental findings from different populations. For example, when studying the protective effects of a vaccine, we may expect the baseline prevalence of the disease in question to differ across populations but the vaccine efficacy to be approximately constant. Our approach can seamlessly encode this belief and naturally admit informative priors on treatment effects derived from external studies, thereby leveraging accumulated knowledge to better identify the effects of treatment.

Similarly, we can account for stratified treatment effects by incorporating discrete covariates into the model, which is equivalent to analyzing a trial within each stratum. Partial pooling would then allow for sharing of information about treatment effects across strata. Another possibility is to tailor our method to the analysis of crossover trials. Under certain assumptions of temporal homogeneity, the efficacy and side effects may be identifiable, making our parameterization and prior proposal natural candidates to the study of treatment effects in such designs. Other relevant extensions include: accounting for confounding and noncompliance; allowing for continuous covariates; and handling adaptive trial designs. With the growing use of Bayesian methods in drug development and the evaluation of clin-

ical trials, these would be timely contributions. More generally, working on the BREASE methodology has elucidated important aspects of statistical and causal inference that merit further study, including the use of Bayes factors in hypothesis testing and sensitivity analysis of treatment effects in a Bayesian context.

BIBLIOGRAPHY

Abouk, R., & Heydari, B. (2021). The immediate effect of COVID-19 policies on social-distancing behavior in the United States. *Public Health Reports*, *136*(2), 245–252.

Acemoglu, D., Chernozhukov, V., Werning, I., & Whinston, M. D. (2021). Optimal Targeted Lockdowns in a Multigroup SIR Model. *American Economic Review: Insights*, *3*(4), 487–502. https://doi.org/10.1257/aeri.20200590

Adda, J. (2016). Economic Activity and the Spread of Viral Diseases: Evidence from High Frequency Data. *The Quarterly Journal of Economics*, *131*(2), 891–941. https://doi.org/10.1093/qje/qjw005

Adolph, C., Amano, K., Bang-Jensen, B., Fullman, N., Magistro, B., Reinke, G., Castellano, R., Erickson, M., & Wilkerson, J. (2022). The pandemic policy U-turn: Partisanship, public health, and race in decisions to ease COVID-19 social distancing policies in the United States. *Perspectives on Politics*, *20*(2), 595–617.

Adolph, C., Amano, K., Bang-Jensen, B., Fullman, N., Magistro, B., Reinke, G., & Wilkerson, J. (2022). Governor partisanship explains the adoption of statewide mask mandates in response to COVID-19. *State Politics & Policy Quarterly*, *22*(1), 24–49.

Adolph, C., Amano, K., Bang-Jensen, B., Fullman, N., & Wilkerson, J. (2021). Pandemic politics: Timing state-level social distancing responses to COVID-19. *Journal of Health Politics, Policy and Law*, *46*(2), 211–233.

Agresti, A., & Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, *14*(3), 297–330.

Agresti, A., & Min, Y. (2005). Frequentist Performance of Bayesian Confidence Intervals for Comparing Proportions in 2 × 2 Contingency Tables. *Biometrics*, *61*(2), 515–523.

Albert, J. H., & Gupta, A. K. (1982). Mixtures of Dirichlet Distributions and Estimation in Contingency Tables. *The Annals of Statistics*, *10*(4), 1261–1268.

Albert, J. H., & Gupta, A. K. (1983a). Bayesian Estimation Methods for 2 × 2 Contingency Tables Using Mixtures of Dirichlet Distributions. *Journal of the American Statistical Association*, *78*(383), 708–717.

Albert, J. H., & Gupta, A. K. (1983b). Estimation in contingency tables using prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, *45*(1), 60–69.

Albert, J. H., & Gupta, A. K. (1985). Bayesian methods for binomial data with applications to a nonresponse problem. *J. Amer. Statist. Assoc.*, *80*, 167–174.

Alexander, D., & Karger, E. (2023). Do Stay-at-Home Orders Cause People to Stay at Home? Effects of Stay-at-Home Orders on Consumer Behavior. *The Review of Economics and Statistics*, *105*(4), 1017–1027. https://doi.org/10.1162/rest_a_01108

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, *191*, 104254. https://doi.org/https://doi.org/10.1016/j.jpubeco.2020.104254

Alvarez, F., Argente, D., & Lippi, F. (2021). A Simple Planning Problem for COVID-19 Lock-down, Testing, and Tracing. *American Economic Review: Insights*, *3*(3), 367–82. https://doi.org/10.1257/aeri.20200201

Antelman, G. R. (1972). Interrelated Bernoulli Processes. *Journal of the American Statistical Association*, *67*(340), 831–841.

Arons, M. M., Hatfield, K. M., Reddy, S. C., Kimball, A., James, A., Jacobs, J. R., Taylor, J., Spicer, K., Bardossy, A. C., Oakley, L. P., et al. (2020). Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *New England journal of medicine*, *382*(22), 2081–2090.

Auger, K. A., Shah, S. S., Richardson, T., Hartley, D., Hall, M., Warniment, A., Timmons, K., Bosse, D., Ferris, S. A., Brady, P. W., et al. (2020). Association between statewide school closure and COVID-19 incidence and mortality in the US. *Jama*, *324*(9), 859–870.

Aum, S., Lee, S. Y. (, & Shin, Y. (2021). COVID-19 doesn't need lockdowns to destroy jobs: The effect of local outbreaks in Korea. *Labour Economics*, *70*, 101993. https://doi.org/https://doi.org/10.1016/j.labeco.2021.101993

Azevedo, J. P., Hasan, A., Goldemberg, D., Geven, K., & Iqbal, S. A. (2021). Simulating the Potential Impacts of COVID-19 School Closures on Schooling and Learning Outcomes: A Set of Global Estimates. *The World Bank Research Observer*, *36*(1), 1–40. https://doi.org/10.1093/wbro/lkab003

Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., & Gardner, L. M. (2020). Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases*, *20*(11), 1247–1254.

Baek, C., McCrory, P. B., Messer, T., & Mui, P. (2021). Unemployment Effects of Stay-at-Home Orders: Evidence from High-Frequency Claims Data. *The Review of Economics and Statistics*, *103*(5), 979–993. https://doi.org/10.1162/rest_a_00996

Baker, S. R., Farrokhnia, R. A., Meyer, S., Pagel, M., & Yannelis, C. (2020). How Does Household Spending Respond to an Epidemic? Consumption during the 2020 COVID-19 Pandemic. *The Review of Asset Pricing Studies*, *10*(4), 834–862. https://doi.org/10.1093/rapstu/raaa009

Banholzer, N., Van Weenen, E., Lison, A., Cenedese, A., Seeliger, A., Kratzwald, B., Tschernutter, D., Salles, J. P., Bottrighi, P., & Lehtinen, S. e. a. (2021). Estimating the effects of non-pharmaceutical interventions on the number of new infections with COVID-19 during the first epidemic wave. *PLoS one*, *16*(6), e0252827.

Barnett-Howell, Z., Watson, O. J., & Mobarak, A. M. (2021). The benefits and costs of social distancing in high- and low-income countries. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, *115*(7), 807–819. https://doi.org/10.1093/trstmh/traa140

Barrios, J. M., & Hochberg, Y. V. (2021). Risk perceptions and politics: Evidence from the COVID-19 pandemic. *Journal of Financial Economics*, *142*(2), 862–879. https://doi.org/https://doi.org/10.1016/j.jfineco.2021.05.039

Barro, R. J., Ursúa, J. F., & Weng, J. (2020). *The coronavirus and the great influenza pandemic: Lessons from the "spanish flu" for the coronavirus's potential effects on mortality and economic activity* (tech. rep.). National Bureau of Economic Research.

Barrot, J.-N., Bonelli, M., Grassi, B., & Sauvagnat, J. (2024). Causal effects of closing businesses in a pandemic. *Journal of Financial Economics*, *154*, 103794. https://doi.org/https://doi.org/10.1016/j.jfineco.2024.103794

Bartik, A. W., Bertrand, M., Lin, F., Rothstein, J., & Unrath, M. (2020). Measuring the Labor Market at the Onset of the COVID-19 Crisis. *Brookings Papers on Economic Activity*.

Bartsch, S. M., Ferguson, M. C., McKinnell, J. A., O'shea, K. J., Wedlock, P. T., Siegmund, S. S., & Lee, B. Y. (2020). The Potential Health Care Costs And Resource Use Associated With COVID-19 In The United States: A simulation estimate of the direct medical costs and health care resource use associated with COVID-19 infections in the United States. *Health affairs*, *39*(6), 927–935.

Bartsch, S. M., O'Shea, K. J., Chin, K. L., Strych, U., Ferguson, M. C., Bottazzi, M. E., Wedlock, P. T., Cox, S. N., Siegmund, S. S., Hotez, P. J., et al. (2022). Maintaining face mask use before and after achieving different COVID-19 vaccination coverage levels: a modelling study. *The Lancet Public Health*, *7*(4), e356–e365.

Basu, D., & Pereira, C. A. d. B. (1982). On the Bayesian analysis of categorical data: the problem of nonresponse. *J. Statist. Plann. Inference*, *6*, 345–362.

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances, with Richard Price's foreword and discussion. *Philos. Trans. R. Soc. London*, *53*, 370–418.

Bayham, J., & Fenichel, E. P. (2020). Impact of school closures for COVID-19 on the US health-care workforce and net mortality: a modelling study. *The Lancet Public Health*, *5*(5), e271–e278.

Benatia, D., Godefroy, R., & Lewis, J. (2020). Estimating COVID-19 prevalence in the United States: a sample selection model approach. *medRxiv*, 2020–04.

Betthäuser, B. A., Bach-Mortensen, A. M., & Engzell, P. (2023). A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. *Nature Human Behaviour*, *7*(3), 375–385.

Blackburn, J., Yiannoutsos, C. T., Carroll, A. E., Halverson, P. K., & Menachemi, N. (2021). Infection fatality ratios for COVID-19 among noninstitutionalized persons 12 and older: results of a random-sample prevalence study. *Annals of Internal Medicine*, *174*(1), 135–136.

Bloom, D. E., Kuhn, M., & Prettner, K. (2022). Modern Infectious Diseases: Macroeconomic Impacts and Policy Responses. *Journal of Economic Literature*, *60*(1), 85–131. https://doi.org/10.1257/jel.20201642

Bo, Y., Guo, C., Lin, C., Zeng, Y., Li, H. B., Zhang, Y., Hossain, M. S., Chan, J. W., Yeung, D. W., Kwok, K. O., et al. (2021). Effectiveness of non-pharmaceutical interventions on COVID-19 transmission in 190 countries from 23 January to 13 April 2020. *International Journal of Infectious Diseases*, *102*, 247–253.

Bodenstein, M., Corsetti, G., & Guerrieri, L. (2022). Social distancing and supply disruptions in a pandemic. *Quantitative Economics*, *13*(2), 681–721.

Böger, B., Fachi, M. M., Vilhena, R. O., Cobre, A. F., Tonin, F. S., & Pontarolo, R. (2021). Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *American Journal of Infection Control*, *49*(1), 21–29.

Branscum, A., Gardner, I., & Johnson, W. (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive veterinary medicine*, *68*(2-4), 145–163.

Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A. B., Leech, G., Altman, G., Mikulik, V., et al. (2021). Inferring the effectiveness of government interventions against COVID-19. *Science*, *371*(6531), eabd9338.

Brodeur, A., Gray, D., Islam, A., & Bhuiyan, S. (2021). A literature review of the economics of COVID-19. *Journal of Economic Surveys*, *35*(4), 1007–1044. https://doi.org/https://doi.org/10.1111/joes.12423

Brown, S. T., Tai, J. H., Bailey, R. R., Cooley, P. C., Wheaton, W. D., Potter, M. A., Voorhees, R. E., LeJeune, M., Grefenstette, J. J., Burke, D. S., et al. (2011). Would school closure for the 2009 H1N1 influenza epidemic have been worth the cost?: a computational simulation of Pennsylvania. *BMC public health*, *11*, 1–11.

Bruns, R., & Teran, N. (2022). *Weighing the Cost of the Pandemic: Knowing what we know now, how much damage did COVID-19 cause in the United States*. https://ifp.org/weighing-the-cost-of-the-pandemic/ Accessed 2024/05/19.

Brzezinski, A., Kecht, V., & Van Dijcke, D. (2020). The Cost of Staying Open: Voluntary Social Distancing and Lockdowns in the US. *Economics Series Working Papers, University of Oxford, Department of Economics*, (910). https://doi.org/10.2139/ssrn.3614494

Bullard, J., Dust, K., Funk, D., Strong, J. E., Alexander, D., Garnett, L., Boodman, C., Bello, A., Hedley, A., Schiffman, Z., et al. (2020). Predicting infectious severe acute

respiratory syndrome coronavirus 2 from diagnostic samples. *Clinical infectious diseases*, *71*(10), 2663–2666.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, *80*, 1–28.

Byrne, A. W., McEvoy, D., Collins, A. B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E. A., McAloon, C., et al. (2020). Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ open*, *10*(8), e039856.

Campbell, H., De Valpine, P., Maxwell, L., De Jong, V. M., Debray, T. P., Jaenisch, T., & Gustafson, P. (2022). Bayesian adjustment for preferential testing in estimating infection fatality rates, as motivated by the COVID-19 pandemic. *The Annals of Applied Statistics*, *16*(1), 436–459.

Campbell, H., & Gustafson, P. (2022). Bayes factors and posterior estimation: Two sides of the very same coin. *The American Statistician*, *0*(0), 1–11.

Casella, G., & Moreno, E. (2009). Assessing robustness of intrinsic tests of independence in two-way contingency tables. *Journal of the American Statistical Association*, *104*(487), 1261–1271.

Cauchemez, S., Valleron, A.-J., Boelle, P.-Y., Flahault, A., & Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, *452*(7188), 750–754.

CDC. (2020). *Duration of isolation and precautions for adults with COVID-19* [Accessed 2021/01/01]. https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html

CDC. (2021). *Estimated Disease Burden of COVID-19* [Accessed 2021/03/01.]. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html

Chen, S., Igan, D., Pierri, N., & Presbitero, A. F. (2020). Tracking the Economic Impact of COVID-19 and Mitigation Policies in Europe and the United States. *IMF Working Papers*, (2020/125). https://www.imf.org/en/Publications/WP/Issues/2020/07/10/Tracking-the-Economic-Impact-of-COVID-19-and-Mitigation-Policies-in-Europe-and-the-United-49553

Chernozhukov, V., Kasahara, H., & Schrimpf, P. (2021). Causal impact of masks, policies, behavior on early COVID-19 pandemic in the U.S. *Journal of Econometrics*, *220*(1), 23–62. https://doi.org/10.1016/j.jeconom.2020.09.003

Chetty, R., Friedman, J. N., & Stepner, M. (2024). The economic impacts of COVID-19: Evidence from a new public database built using private sector data. *The Quarterly Journal of Economics*, *139*(2), 829–889.

Chickering, D. M., & Pearl, J. (1996). A clinician's tool for analyzing non-compliance. *Proceedings of the AAAI Conference on Artificial Intelligence, 13*.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., y Piontti, A. P., Mu, K., Rossi, L., Sun, K., Viboud, C., Xiong, X., Yu, H., Halloran, M. E., Longini, I. M., & Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, *368*(6489), 395–400. https://doi.org/10.1126/science.aba9757

Christakis, D. A., Van Cleve, W., & Zimmerman, F. J. (2020). Estimation of US Children's Educational Attainment and Years of Life Lost Associated With Primary School Closures During the Coronavirus Disease 2019 Pandemic. *JAMA Network Open*, *3*(11), e2028786–e2028786. https://doi.org/10.1001/jamanetworkopen.2020.28786

Cinelli, C., & Pearl, J. (2021). Generalizing experimental results by leveraging knowledge of mechanisms. *European Journal of Epidemiology*, *36*, 149–164.

Coibion, O., Gorodnichenko, Y., & Weber, M. (2020). *The Cost of the COVID-19 Crisis: Lockdowns, Macroeconomic Expectations, and Consumer Spending* (Working Paper No. 27141). National Bureau of Economic Research. https://doi.org/10.3386/w27141

Congressional Budget Office. (2020, September). *An Update to the Budget Outlook: 2020 to 2030*. https://www.cbo.gov/publication/56517 Accessed 2021/09/25.

Copas, J. B. (1973). Randomization models for the matched and unmatched 2 × 2 tables. *Biometrika*, *60*(3), 467–476. Retrieved October 24, 2023, from http://www.jstor.org/stable/2334995

Craig, B. A., & Black, M. A. (2001). Incremental cost-effectiveness ratio and incremental net-health benefit: two sides of the same coin. *Expert Review of Pharmacoeconomics & Outcomes Research*, *1*(1), 37–46.

Cronin, C. J., & Evans, W. N. (2020). *Private precaution and public restrictions: what drives social distancing and industry foot traffic in the COVID-19 era?* (Tech. rep.). National Bureau of Economic Research.

Crucini, M. J., & O'Flaherty, O. (2020). *Stay-at-Home Orders in a Fiscal Union* (Working Paper No. 28182). National Bureau of Economic Research. https://doi.org/10.3386/w28182

Cutler, D. M., & Summers, L. H. (2020). The COVID-19 pandemic and the $16 trillion virus. *Jama*, *324*(15), 1495–1496.

Dablander, F., Huth, K., Gronau, Q. F., Etz, A., & Wagenmakers, E.-J. (2022). A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions. *Statistics in Medicine*, *41*(8), 1319–1333.

Dauelsberg, L. R., Maskery, B., Joo, H., Germann, T. C., Del Valle, S. Y., & Uzicanin, A. (2024). Cost effectiveness of preemptive school closures to mitigate pandemic influenza outbreaks of differing severity in the United States. *BMC Public Health*, *24*(1), 200.

Davies, H., Crombie, I., & Tavakoli, M. (1998). When can odds ratios mislead? *BMJ*, *316*(7136), 989–991.

Davis, E. L., Lucas, T. C., Borlase, A., Pollington, T. M., Abbott, S., Ayabina, D., Crellen, T., Hellewell, J., Pi, L., et al. (2021). Contact tracing is an imperfect tool for controlling COVID-19 transmission and relies on population adherence. *Nature communications*, *12*(1), 5412.

Deb, P., Furceri, D., Ostry, J. D., & Tawk, N. (2022). The economic effects of COVID-19 containment measures. *Open Economies Review*, *33*(1), 1–32.

Decerf, B., Ferreira, F. H., Mahler, D. G., & Sterck, O. (2021). Lives and livelihoods: estimates of the global mortality and poverty effects of the COVID-19 pandemic. *World Development*, *146*, 105561.

DeMartino, J. K., Swallow, E., Goldschmidt, D., Yang, K., Viola, M., Radtke, T., & Kirson, N. (2022). Direct health care costs associated with COVID-19 in the United States [PMID: 35722829]. *Journal of Managed Care & Specialty Pharmacy*, *28*(9), 936–947. https://doi.org/10.18553/jmcp.2022.22050

Demirgüç-Kunt, A., Lokshin, M., & Torre, I. (2021). The sooner, the better: the economic impact of non-pharmaceutical interventions during the early stage of the COVID-19 pandemic. *Economics of Transition and Institutional Change*, *29*(4), 551–573.

Dickey, J. M., Jiang, J. M., & Kadane, J. B. (1987). Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, *82*, 773–781.

Dickey, J. M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, *78*(383), 628–637.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226.

Ding, P., & Miratrix, L. W. (2019). Model-free causal inference of binary experimental data. *Scandinavian Journal of Statistics*, *46*(1), 200–214.

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd, pp. 593–630). Bingley: Emerald.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, *20*(5), 533–534.

Eales, O., Haw, D., Wang, H., Atchison, C., Ashby, D., Cooke, G. S., Barclay, W., Ward, H., Darzi, A., Donnelly, C. A., Chadeau-Hyam, M., Elliott, P., & Riley, S. (2023). Dynamics of SARS-CoV-2 infection hospitalisation and infection fatality ratios over 23 months in England. *PLOS Biology*, *21*(5), 1–21. https://doi.org/10.1371/journal.pbio.3002118

Eichenbaum, M. S., Rebelo, S., & Trabandt, M. (2021). The Macroeconomics of Epidemics. *The Review of Financial Studies*, *34*(11), 5149–5187. https://doi.org/10.1093/rfs/hhab040

Ellis, P. (2020). *Test positivity rates and disease incidence* [Accessed 2021/01/01]. http://freerangestats.info/blog/2020/05/09/covid-population-incidence

Endo, A., Abbott, S., Kucharski, A. J., Funk, S., et al. (2020). Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research*, *5*.

Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, *1*(4), 893–914.

Fahle, E. M., Kane, T. J., Patterson, T., Reardon, S. F., Staiger, D. O., & Stuart, E. A. (2023). School district and community factors associated with learning loss during the COVID-19 pandemic. *Center for Education Policy Research at Harvard University: Cambridge, MA, USA*.

Farboodi, M., Jarosch, G., & Shimer, R. (2021). Internal and external effects of social distancing in a pandemic. *Journal of Economic Theory*, *196*, 105293. https://doi.org/https://doi.org/10.1016/j.jet.2021.105293

Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature*, *442*(7101), 448–452.

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College COVID-19 Response Team. *Imperial College COVID-19 Response Team*, *20*(10.25561), 77482.

Fernández-Villaverde, J., & Jones, C. I. (2020). *Macroeconomic outcomes and COVID-19: a progress report* (tech. rep.). National Bureau of Economic Research.

Ferwana, I., & Varshney, L. R. (2024). The impact of COVID-19 lockdowns on mental health patient populations in the United States. *Scientific reports*, *14*(1), 5689.

Fields, V. L., Kracalik, I. T., Carthel, C., Lopez, A., Schwartz, A., Lewis, N. M., Bray, M., Claflin, C., Jorgensen, K., Khong, H., et al. (2021). Coronavirus Disease Contact Tracing Outcomes and Cost, Salt Lake County, Utah, USA, March–May 2020. *Emerging Infectious Diseases*, *27*(12), 2999.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, *584*(7820), 257–261.

Forsythe, E., Kahn, L. B., Lange, F., & Wiczer, D. (2020). Labor demand in the time of COVID-19: Evidence from vacancy postings and UI claims. *Journal of public economics*, *189*, 104238.

Fuchs-Schündeln, N., Krueger, D., Ludwig, A., & Popova, I. (2022). The Long-Term Distributional and Welfare Effects of COVID-19 School Closures. *The Economic Journal*, *132*(645), 1647–1683. https://doi.org/10.1093/ej/ueac028

Gabry, J., Češnovar, R., & Johnson, A. (2024). *cmdstanr: R Interface to 'CmdStan'*. https://mc-stan.org/cmdstanr

Gallo, L. G., Oliveira, A. F. d. M., Abrahao, A. A., Sandoval, L. A. M., Martins, Y. R. A., Almirón, M., Dos Santos, F. S. G., Araujo, W. N., de Oliveira, M. R. F., & Peixoto, H. M. (2020). Ten epidemiological parameters of COVID-19: use of rapid literature review to inform predictive models during the pandemic. *Frontiers in public health*, *8*, 598547.

Gavenčiak, T., Monrad, J. T., Leech, G., Sharma, M., Mindermann, S., Bhatt, S., Brauner, J., & Kulveit, J. (2022). Seasonal variation in SARS-CoV-2 transmission in temperate climates: A Bayesian modelling study in 143 European regions. *PLoS computational biology*, *18*(8), e1010435.

Ge, Y., Zhang, W.-B., Wu, X., Ruktanonchai, C. W., Liu, H., Wang, J., Song, Y., Liu, M., Yan, W., Yang, J., et al. (2022). Untangling the changing impact of non-pharmaceutical interventions and vaccination on European COVID-19 trajectories. *Nature Communications*, *13*(1), 3106.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman; Hall/CRC.

Gerber, F., & Furrer, R. (2019). optimParallel: An R Package Providing a Parallel Version of the L-BFGS-B Optimization Method. *The R Journal*, *11*(1), 352–358. https://doi.org/10.32614/RJ-2019-030

Gollier, C. (2020). Cost–benefit analysis of age-specific deconfinement strategies. *Journal of Public Economic Theory*, *22*(6), 1746–1771.

Goolsbee, A., & Syverson, C. (2021). Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020. *Journal of Public Economics*, *193*, 104311. https://doi.org/https://doi.org/10.1016/j.jpubeco.2020.104311

Greenhalgh, T., MacIntyre, C. R., Baker, M. G., Bhattacharjee, S., Chughtai, A. A., Fisman, D., Kunasekaran, M., Kvalsvig, A., Lupton, D., Oliver, M., et al. (2024). Masks and respirators for prevention of respiratory infections: a state of the science review. *Clinical microbiology reviews*, e00124–23.

Greenland, S., & Robins, J. (1986). Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, *15*(3), 413–419.

Greenstone, M., & Nigam, V. (2020). Does social distancing matter? *University of Chicago, Becker Friedman Institute for economics working paper*, (2020-26).

Gronau, Q. F., Raj, K. N. A., & Wagenmakers, E.-J. (2021). Informed Bayesian Inference for the A/B Test. *Journal of Statistical Software*, *100*(17), 1–39.

Gu, Y. (2021). *Estimating true infections: A simple heuristic to measure implied infection fatality rate* [Accessed 2021/03/01]. https://covid19-projections.com/estimating-true-infections/

Gunel, E. (1984). A Bayesian analysis of the multinomial model for a dichotomous response with nonrespondents. *Comm. Statist. Theory Methods*, *13*, 737–751.

Gunel, E., & Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika*, *61*(3), 545–557.

Gupta, S., Montenovo, L., Nguyen, T., Lozano-Rojas, F., Schmutte, I., Simon, K., Weinberg, B. A., & Wing, C. (2023). Effects of social distancing policy on labor market outcomes. *Contemporary Economic Policy*, *41*(1), 166–193. https://doi.org/https://doi.org/10.1111/coep.12582

Gustafson, P. (2015). *Bayesian inference for partially identified models: Exploring the limits of limited data* (Vol. 140). CRC Press.

Halder, N., Kelso, J. K., & Milne, G. J. (2011). Cost-effective strategies for mitigating a future influenza pandemic with H1N1 2009 characteristics. *PLoS One*, *6*(7), e22087.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01079-8

Hall, R. E., Jones, C. I., & Klenow, P. J. (2020). *Trading off consumption and COVID-19 deaths* (tech. rep.). National Bureau of Economic Research.

Hall, V., Foulkes, S., Insalata, F., Kirwan, P., Saei, A., Atti, A., Wellington, E., Khawam, J., Munro, K., Cole, M., et al. (2022). Protection against SARS-CoV-2 after Covid-19 vaccination and previous infection. *New England Journal of Medicine*, *386*(13), 1207–1220.

Halloran, M. E., Ferguson, N. M., Eubank, S., Longini Jr, I. M., Cummings, D. A., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T. C., et al. (2008). Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences*, *105*(12), 4639–4644.

Hanushek, E. A., & Woessmann, L. (2020). The economic impacts of learning losses. (225). https://doi.org/https://doi.org/https://doi.org/10.1787/21908d74-en

Hasan, A., Susanto, H., Kasim, M. F., Nuraini, N., Lestari, B., Triany, D., & Widyastuti, W. (2020). Superspreading in early transmissions of COVID-19 in Indonesia. *Scientific reports*, *10*(1), 22386.

Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars-Larrive, A., Loreto, V., Pinior, B., Thurner, S., & Klimek, P. (2020). Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature human behaviour*, *4*(12), 1303–1312.

Havers, F. P., Reed, C., Lim, T., Montgomery, J. M., Klena, J. D., Hall, A. J., Fry, A. M., Cannon, D. L., Chiang, C.-F., Gibbons, A., et al. (2020). Seroprevalence of antibodies

to SARS-CoV-2 in 10 sites in the United States, March 23-May 12, 2020. *JAMA internal medicine*, *180*(12), 1576–1586.

He, J.-L., Luo, L., Luo, Z.-D., Lyu, J.-X., Ng, M.-Y., Shen, X.-P., & Wen, Z. (2020). Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China. *Respiratory medicine*, *168*, 105980.

He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine*, *26*(5), 672–675.

Helfand, M., Fiordalisi, C., Wiedrick, J., Ramsey, K. L., Armstrong, C., Gean, E., Winchell, K., & Arkhipova-Jenkins, I. (2022). Risk for reinfection after SARS-CoV-2: a living, rapid review for American College of Physicians Practice Points on the role of the antibody response in conferring immunity following SARS-CoV-2 infection. *Annals of internal medicine*, *175*(4), 547–555.

Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., Munday, J. D., Kucharski, A. J., Edmunds, W. J., Sun, F., et al. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, *8*(4), e488–e496.

Henry, T. A. (2020). *Experts: Here's how many more contact tracers U.S. needs* [Accessed 2024/06/05.]. https://www.ama-assn.org/delivering-care/public-health/experts-here-s-how-many-more-contact-tracers-us-needs

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*(1), 69–88.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Holbrook, A. L., Krosnick, J. A., & Pfent, A. (2007). The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms. *Advances in Telephone Survey Methodology*, 499–528. https://doi.org/10.1002/9780470173404.ch23

Holtz, D., Zhao, M., Benzell, S. G., Cao, C. Y., Rahimian, M. A., Yang, J., Allen, J., Collis, A., Moehring, A., Sowrirajan, T., et al. (2020). Interdependence and the cost of uncoordinated responses to COVID-19. *Proceedings of the National Academy of Sciences*, *117*(33), 19837–19843.

Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., Druckenmiller, H., Huang, L. Y., Hultgren, A., Krasovich, E., et al. (2020). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*, *584*(7820), 262–267.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, *25*(1), 305–327.

Irons, N. J., & Cinelli, C. (2023). Causally sound priors for binary experiments. *arXiv preprint arXiv:2308.13713*.

Irons, N. J., & Raftery, A. E. (2021). Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proceedings of the National Academy of Sciences*, *118*(31), e2103272118.

Jamison, J. C., Bundy, D., Jamison, D. T., Spitz, J., & Verguet, S. (2021). Comparing the impact on COVID-19 mortality of self-imposed behavior change and of government regulations across 13 countries. *Health services research*, *56*(5), 874–884.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(2), 203–222.

Jeffreys, H. (1961). *Theory of Probability* (3rd). Oxford, UK: Oxford University Press.

Jin, R. (2021). The lag between daily reported COVID-19 cases and deaths and its relationship to age. *Journal of public health research*, *10*(3), jphr–2021.

Johndrow, J., Ball, P., Gargiulo, M., & Lum, K. (2020). Estimating the number of SARS-CoV-2 infections and the impact of mitigation policies in the United States. *Harvard Data Sci. Rev*, *10*.

Jones, C., Philippon, T., & Venkateswaran, V. (2021). Optimal Mitigation Policies in a Pandemic: Social Distancing and Working from Home. *The Review of Financial Studies*, *34*(11), 5188–5223. https://doi.org/10.1093/rfs/hhab076

Juneau, C.-E., Pueyo, T., Bell, M., Gee, G., Collazzo, P., & Potvin, L. (2022). Lessons from past pandemics: a systematic review of evidence-based, cost-effective interventions to suppress COVID-19. *Systematic reviews*, *11*(1), 90.

Kaplan, S., Lefler, J., & Zilberman, D. (2022). The political economy of COVID-19. *Applied Economic Perspectives and Policy*, *44*(1), 477–488. https://doi.org/https://doi.org/10.1002/aepp.13164

Karaivanov, A., Lu, S. E., Shigeoka, H., Chen, C., & Pamplona, S. (2021). Face masks, public policies and slowing the spread of COVID-19: Evidence from Canada. *Journal of Health Economics*, *78*, 102475. https://doi.org/https://doi.org/10.1016/j.jhealeco.2021.102475

Karson, M., & Wrobleski, W. (1970). A Bayesian Analysis of a Binomial Model with a Partially Informative Category. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 532–534.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *54*(1), 129–144.

Kass, R. E., & Wasserman, L. (1995). A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, *90*(431), 928–934.

Katz, J., Lu, D., & Sanger-Katz, M. (2021). 494,000 More U.S. Deaths Than Normal Since Covid-19 Struck. *The New York Times*. https://www.nytimes.com/interactive/2021/01/14/us/covid-19-death-toll.html

Kaufman, G. M., & King, B. (1973). A Bayesian Analysis of Nonresponse in Dichotomous Processes. *Journal of the American Statistical Association*, *68*(343), 670–678.

Kelso, J. K., Halder, N., Postma, M. J., & Milne, G. J. (2013). Economic analysis of pandemic influenza mitigation strategies for five pandemic severity categories. *BMC public health*, *13*, 1–17.

Keogh-Brown, M. R., Jensen, H. T., Edmunds, W. J., & Smith, R. D. (2020). The impact of Covid-19, associated behaviours and policies on the UK economy: A computable general equilibrium model. *SSM-population health*, *12*, 100651.

Keogh-Brown, M. R., Smith, R. D., Edmunds, J. W., & Beutels, P. (2010). The macroeconomic impact of pandemic influenza: estimates from models of the United Kingdom, France, Belgium and The Netherlands. *The European Journal of Health Economics*, *11*, 543–554.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, *115*(772), 700–721.

Killion, R. A., & Zahn, D. A. (1976). A bibliography of contingency table literature: 1900 to 1974. *International Statistical Review*, *44*(1), 71–112.

Kissane, E., & Rivera, J. M. (2020). *Test positivity in the US is a mess* [Accessed 2021/01/01]. https://covidtracking.com/analysis-updates/test-positivity-in-the-us-is-a-mess

Kline, D., Li, Z., Chu, Y., Wakefield, J., Miller, W. C., Norris Turner, A., & Clark, S. J. (2021). Estimating seroprevalence of SARS-CoV-2 in Ohio: A Bayesian multilevel poststratification approach with multiple diagnostic tests. *Proceedings of the National Academy of Sciences*, *118*(26), e2023947118.

Korea Centers for Disease Control and Prevention. (2020). *Findings from investigation and analysis of re-positive cases* [Accessed 2021/01/01]. https://www.cdc.go.kr/board/board.es?mid=a30402000000%5C&bid=0030%5C&act=view%5C&list_no=367267%5C&nPage=1

Kremer, C., Torneri, A., Boesmans, S., Meuwissen, H., Verdonschot, S., Vanden Driessche, K., Althaus, C. L., Faes, C., & Hens, N. (2021). Quantifying superspreading for COVID-19 using Poisson mixture distributions. *Scientific reports*, *11*(1), 14107.

Krueger, D., Uhlig, H., & Xie, T. (2022). Macroeconomic dynamics and reallocation in an epidemic: evaluating the 'Swedish solution'. *Economic Policy*, *37*(110), 341–398. https://doi.org/10.1093/epolic/eiac010

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Elsevier Science & Technology.

Laplace, P. S. (1774). Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*.

Lash, R. R., Moonan, P. K., Byers, B. L., Bonacci, R. A., Bonner, K. E., Donahue, M., Donovan, C. V., Grome, H. N., Janssen, J. M., Magleby, R., McLaughlin, H. P., Miller, J. S., Pratt, C. Q., Steinberg, J., Varela, K., Anschuetz, G. L., Cieslak, P. R., Fialkowski, V., Fleischauer, A. T., ... Team, C.-1. C. T. A. (2021). COVID-19 Case Investigation and Contact Tracing in the US, 2020. *JAMA Network Open*, *4*(6), e2115850–e2115850. https://doi.org/10.1001/jamanetworkopen.2021.15850

Lau, M. S. Y., Grenfell, B., Thomas, M., Bryan, M., Nelson, K., & Lopman, B. (2020). Characterizing superspreading events and age-specific infectiousness of SARS-CoV-

2 transmission in Georgia, USA. *Proceedings of the National Academy of Sciences*, *117*(36), 22430–22435. https://doi.org/10.1073/pnas.2011802117

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020a). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*, *172*(9), 577–582.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020b). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, *172*(9), 577–582.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. Wiley.

Lempel, H., Epstein, J. M., & Hammond, R. A. (2009). Economic cost and health care workforce effects of school closures in the US. *PLoS currents*, *1*.

Leonard, T. (1972). Bayesian methods for binomial data. *Biometrika*, *59*(3), 581–589.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *37*(1), 23–37.

Li, H., Yuan, K., Sun, Y.-K., Zheng, Y.-B., Xu, Y.-Y., Su, S.-Z., Zhang, Y.-X., Zhong, Y., Wang, Y.-J., & Tian, S.-S. e. a. (2022). Efficacy and practice of facemask use in general population: a systematic review and meta-analysis. *Translational psychiatry*, *12*(1), 49.

Li, Y., Campbell, H., & Kulkarni, D. e. a. (2021). The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries. *The Lancet Infectious Diseases*, *21*(2), 193–202. https://doi.org/10.1016/S1473-3099(20)30785-4

Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., Yuan, B., Kinoshita, R., & 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. *Journal of clinical medicine*, *9*(2), 538.

Liu, Y., Morgenstern, C., Kelly, J., Lowe, R., & Jit, M. (2021). The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC medicine*, *19*, 1–12.

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, *27*(2), taaa021. https://doi.org/10.1093/jtm/taaa021

Lo, J., Cox, C., Amin, K., Telesford, I., Dawson, L., & Kates, J. (2023). *Health Spending: Prices for COVID-19 testing* [Accessed 2024/05/28.]. https://www.healthsystemtracker.org/brief/prices-for-covid-19-testing/

Lordan, R., FitzGerald, G. A., & Grosser, T. (2020). Reopening schools during COVID-19.

Lu, J., Peng, J., Xiong, Q., Liu, Z., Lin, H., Tan, X., Kang, M., Yuan, R., Zeng, L., Zhou, P., et al. (2020). Clinical, immunological and virological characterization of COVID-19 patients that test re-positive for SARS-CoV-2 by RT-PCR. *EBioMedicine*, *59*.

Lyu, W., & Wehby, G. L. (2020). Community Use Of Face Masks And COVID-19: Evidence From A Natural Experiment Of State Mandates In The US: Study examines impact on COVID-19 growth rates associated with state government mandates requiring face mask use in public. *Health affairs*, *39*(8), 1419–1425.

Mader, S., & Rüttenauer, T. (2022). The effects of non-pharmaceutical interventions on COVID-19 mortality: a generalized synthetic control approach across 169 countries. *Frontiers in Public Health*, *10*, 820642.

Madigan, D. (1999). Bayesian graphical models, intention-to-treat, and the Rubin causal model. *Seventh International Workshop on Artificial Intelligence and Statistics*.

Mahajan, S., Srinivasan, R., Redlich, C. A., Huston, S. K., Anastasio, K. M., Cashman, L., Massey, D. S., Dugan, A., Witters, D., Marlar, J., et al. (2021). Seroprevalence of SARS-CoV-2-specific IgG antibodies among adults living in Connecticut: Post-infection prevalence (PIP) study. *The American Journal of Medicine*, *134*(4), 526–534.

Maharaj, S., & Kleczkowski, A. (2012). Controlling epidemic spread by social distancing: Do it well or not at all. *BMC public health*, *12*, 1–16.

Markel, H., Lipman, H. B., Navarro, J. A., Sloan, A., Michalsen, J. R., Stern, A. M., & Cetron, M. S. (2007). Nonpharmaceutical Interventions Implemented by US Cities During the 1918-1919 Influenza Pandemic. *JAMA*, *298*(6), 644–654. https://doi.org/10.1001/jama.298.6.644

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020). Coronavirus Pandemic (COVID-19) [https://ourworldindata.org/coronavirus]. *Our World in Data*.

McElreath, R. (2020). *Statistical rethinking : A bayesian course with examples in r and stan* (Second edition.). Chapman & Hall/CRC.

Menachemi, N., Yiannoutsos, C. T., Dixon, B. E., Duszynski, T. J., Fadel, W. F., Wools-Kaloustian, K. K., Needleman, N. U., Box, K., Caine, V., Norwood, C., Weaver, L., & Halverson, P. K. (2020). Population point prevalence of SARS-CoV-2 infection based on a statewide random sample – Indiana, April 25–29, 2020. *CDC Morbidity and Mortality Weekly Report*, *69*(29), 960–964. http://dx.doi.org/10.15585/mmwr.mm6929e1

Mervosh, S., Miller, C. C., & Paris, F. (2024). *What the data says about pandemic school closures, four years later* [Accessed 2024/06/20.]. https://www.nytimes.com/2024/03/18/upshot/pandemic-school-closures-data.html

Meteostat. (2022). *Meteostat Python*. https://dev.meteostat.net/python/

Metodiev, M., Perrot-Dockès, M., Ouadah, S., Irons, N. J., & Raftery, A. E. (2024). Easily Computed Marginal Likelihoods from Posterior Simulation Using the THAMES Estimator. *Bayesian Analysis*, *1*(1), 1–28.

Meyerowitz-Katz, G., & Merone, L. (2020). A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates. *International Journal of Infectious Diseases*, *101*, 138–148.

Milne, G. J., Halder, N., & Kelso, J. K. (2013). The cost effectiveness of pandemic influenza interventions: a pandemic severity based analysis. *PloS one*, *8*(4), e61504.

National Academies of Sciences, Engineering, and Medicine. (2020). *Evaluating Data Types: A Guide for Decision Makers using Data to Understand the Extent and Spread of COVID-19*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25826

New York City Health Department. (2021). *NYC Coronavirus Disease 2019 (COVID-19) Data* [Accessed 2021/03/01]. https://github.com/nychealth/coronavirus-data

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. [Translated from the 1923 Polish original and edited by D. M. Dabrowska and T. P. Speed]. *Statistical Science*, *5*(4), 465–472.

Ng, K. W., Tang, M.-L., Tan, M., & Tian, G.-L. (2008). Grouped Dirichlet distribution: A new tool for incomplete categorical data analysis. *Journal of Multivariate Analysis*, *99*(3), 490–509.

Ng, K. W., Tian, G.-L., & Tang, M.-L. (2011). *Dirichlet and related distributions: Theory, methods and applications*. Wiley.

Nichols, G. L., Gillingham, E., Macintyre, H., Vardoulakis, S., Hajat, S., Sarran, C., Amankwaah, D., & Phalkey, R. (2021). Coronavirus seasonality, respiratory infections and weather. *BMC Infectious Diseases*, *21*, 1–15.

Painter, M., & Qiu, T. (2021). Political beliefs affect compliance with government mandates. *Journal of Economic Behavior and Organization*, *185*, 688–701. https://doi.org/https://doi.org/10.1016/j.jebo.2021.03.019

Park, T., & Brown, M. B. (1994). Models for categorical data with nonignorable nonresponse. *J. Amer. Statist. Assoc.*, *89*, 44–52.

Pasquini-Descomps, H., Brender, N., & Maradan, D. (2017). Value for money in H1N1 influenza: a systematic review of the cost-effectiveness of pandemic interventions. *Value in Health*, *20*(6), 819–827.

Paulden, M. (2020). Why it's time to abandon the ICER. *Pharmacoeconomics*, *38*(8), 781–784.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Perlroth, D. J., Glass, R. J., Davey, V. J., Cannon, D., Garber, A. M., & Owens, D. K. (2010). Health outcomes and costs of community mitigation strategies for an influenza pandemic in the United States. *Clinical infectious diseases*, *50*(2), 165–174.

Petherick, A., Goldszmidt, R., Andrade, E. B., Furst, R., Hale, T., Pott, A., & Wood, A. (2021). A worldwide assessment of changes in adherence to COVID-19 protective behaviours and hypothesized pandemic fatigue. *Nature Human Behaviour*, *5*(9), 1145–1160.

Pham-Gia, T., & Turkkan, N. (1998). Distribution of the linear combination of two general beta variables and applications. *Communications in Statistics - Theory and Methods*, *27*(7), 1851–1869.

Plummer, M. (2023). *rjags: Bayesian Graphical Models using MCMC* [R package version 4-14].

Podolsky, M. I., Present, I., Neumann, P. J., & Kim, D. D. (2022). A Systematic Review of Economic Evaluations of COVID-19 Interventions: Considerations of Non-Health Impacts and Distributional Issues. *Value in Health*, *25*(8), 1298–1306. https://doi.org/https://doi.org/10.1016/j.jval.2022.02.003

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., ... Gruber, W. C. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, *383*(27), 2603–2615.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, *108*(504), 1339–1349.

Pooley, N., Abdool Karim, S. S., Combadière, B., Ooi, E. E., Harris, R. C., El Guerche Seblain, C., Kisomi, M., & Shaikh, N. (2023). Durability of vaccine-induced and natural immunity against COVID-19: a narrative review. *Infectious Diseases and Therapy*, *12*(2), 367–387.

Prager, F., Wei, D., & Rose, A. (2017). Total economic consequences of an influenza outbreak in the United States. *Risk Analysis*, *37*(1), 4–19.

Psacharopoulos, G., Collis, V., Patrinos, H. A., & Vegas, E. (2021). The COVID-19 Cost of School Closures in Earnings and Income across the World. *Comparative Education Review*, *65*(2), 271–287. https://doi.org/10.1086/713540

Quicke, K., Gallichotte, E., Sexton, N., Young, M., Janich, A., Gahm, G., Carlton, E. J., Ehrhart, N., & Ebel, G. D. (2020). Longitudinal surveillance for SARS-CoV-2 RNA among asymptomatic staff in five Colorado skilled nursing facilities: epidemiologic, virologic and sequence analysis.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing* [https://www.R-project.org/]. R Foundation for Statistical Computing. Vienna, Austria.

Rader, B., White, L. F., Burns, M. R., Chen, J., Brilliant, J., Cohen, J., Shaman, J., Brilliant, L., Kraemer, M. U., & Hawkins, J. B. e. a. (2021). Mask-wearing and control of SARS-CoV-2 transmission in the USA: a cross-sectional study. *The Lancet Digital Health*, *3*(3), e148–e157.

Rainisch, G., Jeon, S., Pappas, D., Spencer, K. D., Fischer, L. S., Adhikari, B. B., Taylor, M. M., Greening, J., Bradford, Moonan, P. K., Oeltmann, J. E., Kahn, E. B., Washington, M. L., & Meltzer, M. I. (2022). Estimated COVID-19 Cases and Hospitalizations Averted by Case Investigation and Contact Tracing in the US. *JAMA Network Open*, *5*(3), e224042–e224042. https://doi.org/10.1001/jamanetworkopen.2022.4042

Recht, B. (2024). *All models are wrong, but some are dangerous* [Accessed 2024/06/17.]. https://www.argmin.net/p/all-models-are-wrong-but-some-are

Reese, H., Iuliano, A. D., Patel, N. N., Garg, S., Kim, L., Silk, B. J., Hall, A. J., Fry, A., & Reed, C. (2020). Estimated Incidence of Coronavirus Disease 2019 (COVID-19) Illness and Hospitalization – United States, February–September 2020. *Clinical Infectious Diseases*. https://doi.org/10.1093/cid/ciaa1780

Reluga, T. C. (2010). Game theory of social distancing in response to an epidemic. *PLoS computational biology*, *6*(5), e1000793.

Rice, K. L., Miller, G. F., Coronado, F., & Meltzer, M. I. (2020). Estimated Resource Costs for Implementation of CDC's Recommended COVID-19 Mitigation Strategies in Pre-Kindergarten through Grade 12 Public Schools — United States, 2020–21 School Year. *MMWR Morb Mortal Wkly Rep*, *69*, 1917–1921. https://doi.org/10.15585/mmwr.mm6950e1

Richardson, T. S., Evans, R. J., & Robins, J. M. (2011). Transparent parametrizations of models for potential outcomes. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P.

Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.), *Bayesian statistics* (Vol. 9). Oxford University Press.

Robbins, H. E. (1992). An Empirical Bayes Approach to Statistics. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Foundations and basic theory* (pp. 388–394). Springer New York.

Robinson, L. A., Sullivan, R., & Shogren, J. F. (2021). Do the Benefits of COVID-19 Policies Exceed the Costs? Exploring Uncertainties in the Age–VSL Relationship. *Risk Analysis*, *41*(5), 761–770. https://doi.org/https://doi.org/10.1111/risa.13561

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Sadique, M. Z., Adams, E. J., & Edmunds, W. J. (2008). Estimating the costs of school closure for mitigating an influenza pandemic. *BMC public health*, *8*, 1–7.

Sander, B., Nizam, A., Garrison, L. P., Postma, M. J., Halloran, M. E., & Longini, I. M. (2009). Economic Evaluation of Influenza Pandemic Mitigation Strategies in the United States Using a Stochastic Microsimulation Transmission Model. *Value in Health*, *12*(2), 226–233. https://doi.org/https://doi.org/10.1111/j.1524-4733.2008.00437.x

Sharfstein, J. (2024). *Q&A: How much does it cost to get a COVID-19 test? It depends* [Accessed 2024/05/28.]. https://coronavirus.jhu.edu/from-our-experts/q-and-a-how-much-does-it-cost-to-get-a-covid-19-test-it-depends

Sharma, M., Mindermann, S., Rogers-Smith, C., Leech, G., Snodin, B., Ahuja, J., Sandbrink, J. B., Monrad, J. T., Altman, G., Dhaliwal, G., et al. (2021). Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. *Nature communications*, *12*(1), 5820.

Simmons-Duffin, S. (2020). *As states reopen, do they have the workforce they need to stop coronavirus outbreaks?* [Accessed 2024/06/05.]. https://www.npr.org/sections/

health-shots/2020/06/18/879787448/as-states-reopen-do-they-have-the-workforce-they-need-to-stop-coronavirus-outbre

Skarp, J. E., Downey, L. E., Ohrnberger, J. W., Cilloni, L., Hogan, A. B., Sykes, A. L., Wang, S. S., Shah, H. A., Xiao, M., & Hauck, K. (2021). A systematic review of the costs relating to non-pharmaceutical interventions against infectious disease outbreaks. *Applied Health Economics and Health Policy*, *19*, 673–697.

Smith, R. D., Keogh-Brown, M. R., Barnett, T., & Tait, J. (2009). The economy-wide impact of pandemic influenza on the UK: a computable general equilibrium modelling experiment. *BMJ*, *339*.

Smith, S. W., Hauben, M., & Aronson, J. K. (2012). Paradoxical and bidirectional drug effects. *Drug safety*, *35*, 173–189.

Sneppen, K., Nielsen, B. F., Taylor, R. J., & Simonsen, L. (2021). Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. *Proceedings of the National Academy of Sciences*, *118*(14), e2016623118. https://doi.org/10.1073/pnas.2016623118

Spencer, K. D. (2021). COVID-19 case investigation and contact tracing efforts from health departments—United States, June 25–July 24, 2020. *MMWR. Morbidity and Mortality Weekly Report*, *70*.

Springer, M. D., & Thompson, W. E. (1970). The Distribution of Products of Beta, Gamma and Gaussian Random Variables. *SIAM Journal on Applied Mathematics*, *18*(4), 721–737.

Stadlbauer, D., Tan, J., Jiang, K., Hernandez, M. M., Fabre, S., Amanat, F., Teo, C., Arunkumar, G. A., McMahon, M., Capuano, C., et al. (2021). Repeated cross-sectional sero-monitoring of SARS-CoV-2 in New York City. *Nature*, *590*(7844), 146–150.

Stan Development Team. (2020). RStan: The R interface to Stan [R package version 2.21.2]. http://mc-stan.org/

Steering Committee of the Physicians' Health Study Research Group. (1989). Final Report on the Aspirin Component of the Ongoing Physicians' Health Study. *New England Journal of Medicine*, *321*(3), 129–135.

Stinnett, A. A., & Mullahy, J. (1998). Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical decision making*, *18*(2_suppl), S68–S80.

Stokes, J., Turner, A. J., Anselmi, L., Morciano, M., & Hone, T. (2022). The relative effects of non-pharmaceutical interventions on wave one COVID-19 mortality: natural experiment in 130 countries. *BMC Public Health*, *22*(1), 1113.

Sun, K., Wang, W., Gao, L., Wang, Y., Luo, K., Ren, L., Zhan, Z., Chen, X., Zhao, S., Huang, Y., Sun, Q., Liu, Z., Litvinova, M., Vespignani, A., Ajelli, M., Viboud, C., & Yu, H. (2021). Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science*, *371*(6526), eabe2424. https://doi.org/10.1126/science.abe2424

Tai, D. B. G., Shah, A., Doubeni, C. A., Sia, I. G., & Wieland, M. L. (2021). The disproportionate impact of COVID-19 on racial and ethnic minorities in the United States. *Clinical infectious diseases*, *72*(4), 703–706.

Talic, S., Shah, S., Wild, H., Gasevic, D., Maharaj, A., Ademi, Z., Li, X., Xu, W., Mesa-Eguiagaray, I., Rostron, J., et al. (2021). Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. *BMJ*, *375*.

Thall, P. F., Simon, R. M., & Estey, E. H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in medicine*, *14*(4), 357–379.

The Atlantic. (2021). *The COVID Tracking Project* [Accessed 2024/06/01]. covidtracking.com

The New York Times COVID-19 Data Team. (2021). Coronavirus (Covid-19) Data in the United States [Retrieved April 15, 2021, from https://github.com/nytimes/covid-19-data]. *The New York Times*.

The Ohio Channel. (2020). Ohio Governor Mike DeWine - COVID-19 Update — October 1, 2020 [https://www.youtube.com/watch?v=oVCSOlyJ16k. See minutes 22:10–26:40. Accessed 2021/01/01]. *YouTube*.

Thunström, L., Newbold, S. C., Finnoff, D., Ashworth, M., & Shogren, J. F. (2020). The Benefits and Costs of Using Social Distancing to Flatten the Curve for COVID-19. *Journal of Benefit-Cost Analysis*, *11*(2), 179–195. https://doi.org/10.1017/bca.2020.12

Tian, G.-L., Ng, K. W., & Geng, Z. (2003). Bayesian computation for contingency tables with incomplete cell-counts. *Statistica Sinica*, *13*(1), 189–206.

Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, *28*(1), 287–313.

UNICEF et al. (2021). *The state of the global education crisis: a path to recovery: a joint UNESCO, UNICEF and WORLD BANK report*. Paris: UNESCO, cop. 2021.

U.S. Bureau of Economic Analysis. (2020). *Gross Domestic Product, Fourth Quarter and Year 2019 (Third Estimate); Corporate Profits, Fourth Quarter and Year 2019* [Accessed 2024/05/28.]. https://www.bea.gov/news/2020/gross-domestic-product-fourth-quarter-and-year-2019-third-estimate-corporate-profits

U.S. Bureau of Economic Analysis. (2022). *State Personal Income: 2nd Quarter 2022 and Annual 2021* [Accessed 2024/05/28.]. https://www.bea.gov/sites/default/files/2022-09/covid-workbook-ann.xlsx

U.S. Bureau of Economic Analysis. (2023a). *Gross Domestic Product by State, Annual 2017-2022* [Accessed 2024/05/28.]. https://apps.bea.gov/regional/histdata/releases/0923gdpstate/SAGDP.zip

U.S. Bureau of Economic Analysis. (2023b). *Personal Consumption Expenditures by State, 2022* [Accessed 2024/05/28.]. https://apps.bea.gov/regional/zip/SAPCE.zip

U.S. Census Bureau. (2023). *Median Personal Income in the United States [MEPAINUSA646N]* [Retrieved from FRED, Federal Reserve Bank of St. Louis; 2024/05/28.]. https://fred.stlouisfed.org/series/MEPAINUSA646N

U.S. Food and Drug Administration. (2020, October). *Guidance for industry: emergency use authorization for vaccines to prevent COVID-19*. https://www.fda.gov/media/142749/download Accessed 2023/07/12.

van Kampen, J. J., van de Vijver, D. A., Fraaij, P. L., Haagmans, B. L., Lamers, M. M., Okba, N., van den Akker, J. P., Endeman, H., Gommers, D. A., Cornelissen, J. J., et al. (2020). Shedding of infectious virus in hospitalized patients with coronavirus disease-2019 (COVID-19): duration and key determinants. *MedRxiv*, 2020–06.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*, 1413–1432.

Verschuur, J., Koks, E. E., & Hall, J. W. (2021). Global economic impacts of COVID-19 lockdown measures stand out in high-frequency shipping data. *PloS one*, *16*(4), e0248818.

Viner, R., Russell, S., Saulle, R., Croker, H., Stansfield, C., Packer, J., Nicholls, D., Goddings, A.-L., Bonell, C., Hudson, L., Hope, S., Ward, J., Schwalbe, N., Morgan, A., & Minozzi, S. (2022). School Closures During Social Lockdown and Mental Health, Health Behaviors, and Well-being Among Children and Adolescents During the First COVID-19 Wave: A Systematic Review. *JAMA Pediatrics*, *176*(4), 400–409. https://doi.org/10.1001/jamapediatrics.2021.5840

Viner, R. M., Bonell, C., Drake, L., Jourdan, D., Davies, N., Baltag, V., Jerrim, J., Proimos, J., & Darzi, A. (2021). Reopening schools during the COVID-19 pandemic: governments must balance the uncertainty and risks of reopening schools against the clear

harms associated with prolonged closure. *Archives of disease in childhood*, *106*(2), 111–113.

Viner, R. M., Russell, S. J., Croker, H., Packer, J., Ward, J., Stansfield, C., Mytton, O., Bonell, C., & Booy, R. (2020). School closure and management practices during coronavirus outbreaks including COVID-19: a rapid systematic review. *The Lancet Child & Adolescent Health*, *4*(5), 397–404.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Walmsley, T., Rose, A., John, R., Wei, D., Hlávka, J. P., Machado, J., & Byrd, K. (2023). Macroeconomic consequences of the COVID-19 pandemic. *Economic modelling*, *120*, 106147.

Wang, X., Du, Z., James, E., Fox, S. J., Lachmann, M., Meyers, L. A., & Bhavnani, D. (2022). The effectiveness of COVID-19 testing and contact tracing in a US city. *Proceedings of the National Academy of Sciences*, *119*(34), e2200652119.

Ward, T., Fyles, M., Glaser, A., Paton, R. S., Ferguson, W., & Overton, C. E. (2024). The real-time infection hospitalisation and fatality risk across the COVID-19 pandemic in England. *Nature Communications*, *15*(1), 4633.

Ward, T., & Johnsen, A. (2021). Understanding an evolving pandemic: An analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *Plos one*, *16*(10), e0257978.

Weinstein, M. C. (1990). Principles of cost-effective resource allocation in health care organizations. *International journal of technology assessment in health care*, *6*(1), 93–103.

Wiemken, T. L., Khan, F., Puzniak, L., Yang, W., Simmering, J., Polgreen, P., Nguyen, J. L., Jodar, L., & McLaughlin, J. M. (2023). Seasonal trends in COVID-19 cases,

hospitalizations, and mortality in the United States and Europe. *Scientific Reports*, *13*(1), 3886.

Wolf, D. A., Monnat, S. M., Wiemers, E. E., Sun, Y., Zhang, X., Grossman, E. R., & Montez, J. K. (2024). State COVID-19 Policies and Drug Overdose Mortality Among Working-Age Adults in the United States, 2020 [PMID: 38696735]. *American Journal of Public Health*, *114*(7), 714–722. https://doi.org/10.2105/AJPH.2024.307621

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature*, *581*(7809), 465–469.

Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford Jr, J. M., Reingold, A., et al. (2020). Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature communications*, *11*(1), 4507.

Wu, Y., Kang, L., Guo, Z., Liu, J., Liu, M., & Liang, W. (2022). Incubation period of COVID-19 caused by unique SARS-CoV-2 strains: a systematic review and meta-analysis. *JAMA network open*, *5*(8), e2228008–e2228008.

Xie, C., Jiang, L., Huang, G., Pu, H., Gong, B., Lin, H., Ma, S., Chen, X., Long, B., Si, G., et al. (2020). Comparison of different samples for 2019 novel coronavirus detection by nucleic acid amplification tests. *International Journal of Infectious Diseases*, *93*, 264–267.

Xue, Y., Kristiansen, I. S., & de Blasio, B. F. (2012). Dynamic modelling of costs and health consequences of school closure during an influenza pandemic. *BMC public health*, *12*, 1–17.

Yang, W., Kandula, S., Huynh, M., Greene, S. K., Van Wye, G., Li, W., Chan, H. T., McGibbon, E., Yeung, A., Olson, D., et al. (2021). Estimating the infection-fatality

risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis. *The Lancet Infectious Diseases*, *21*(2), 203–212.

Yang, W., Shaff, J., & Shaman, J. (2021). Effectiveness of non-pharmaceutical interventions to contain COVID-19: a case study of the 2020 spring pandemic wave in New York City. *Journal of the Royal Society Interface*, *18*(175), 20200822.

Yiannoutsos, C. T., Halverson, P. K., & Menachemi, N. (2021). Bayesian estimation of SARS-CoV-2 prevalence in Indiana by random testing. *Proceedings of the National Academy of Sciences*, *118*(5). https://doi.org/10.1073/pnas.2013906118

Yu, F., Yan, L., Wang, N., Yang, S., Wang, L., Tang, Y., Gao, G., Wang, S., Ma, C., Xie, R., et al. (2020). Quantitative detection and viral load analysis of SARS-CoV-2 in infected patients. *Clinical Infectious Diseases*, *71*(15), 793–798.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, *395*(10229), 1054–1062.

Appendix A

APPENDIX

A.1 SARS-CoV-2 infection estimates for all US states

Tables A.1 and A.2 list estimates of the COVID infection fatality rate (IFR), cumulative viral incidence, and the infection undercount factor (total infections divided by total cases) for all US states as of January 6, 2021.

Figures 1 through 51 display daily estimates of new infections, cumulative incidence, the SIR reproductive number $r(t) = \beta_t/\gamma$, and the undercount factor for all states from March 2020 to January 2021.

Table A.1: Posterior median and 95% intervals for IFR, cumulative incidence as of March 7, 2021, and undercount factor as of March 7, 2021.

State	IFR	Cumulative Incidence	Undercount
Alabama	0.90% (0.70-1.15)	24.0% (18.8-30.6)	2.4 (1.8-3.0)
Alaska	0.35% (0.28-0.43)	11.8% (10.1-14.2)	1.5 (1.3-1.8)
Arizona	0.93% (0.75-1.14)	24.5% (20.1-30.3)	2.2 (1.8-2.7)
Arkansas	0.78% (0.64-0.95)	23.2% (19.2-28.4)	2.2 (1.8-2.7)
California	0.88% (0.70-1.09)	16.9% (13.6-20.9)	1.9 (1.6-2.4)
Colorado	0.60% (0.49-0.71)	17.2% (14.6-20.9)	2.3 (2.0-2.8)
Connecticut	1.37% (1.10-1.70)	15.9% (12.9-19.9)	2.0 (1.6-2.5)
Delaware	0.92% (0.74-1.13)	17.0% (14.0-21.0)	1.9 (1.6-2.3)
Florida	0.78% (0.64-0.94)	19.5% (16.3-23.8)	2.2 (1.9-2.8)
Georgia	0.67% (0.54-0.82)	25.9% (21.4-32.0)	2.7 (2.2-3.4)
Hawaii	0.70% (0.54-0.91)	4.6% (3.6-5.8)	2.3 (1.8-2.9)
Idaho	0.28% (0.23-0.33)	37.8% (32.1-45.9)	4.0 (3.4-4.8)
Illinois	0.96% (0.78-1.16)	19.3% (16.1-23.7)	2.0 (1.7-2.5)
Indiana	0.84% (0.70-1.00)	22.9% (19.2-27.6)	2.3 (1.9-2.8)
Iowa	0.76% (0.61-0.97)	23.6% (18.7-29.5)	2.6 (2.1-3.3)
Kansas	0.59% (0.47-0.75)	28.6% (22.6-36.1)	2.8 (2.2-3.5)
Kentucky	0.53% (0.43-0.63)	21.4% (18.0-25.8)	2.3 (2.0-2.8)
Louisiana	1.01% (0.82-1.22)	21.1% (17.5-26.1)	2.3 (1.9-2.8)
Maine	0.73% (0.60-0.87)	7.3% (6.2-8.8)	2.1 (1.8-2.6)
Maryland	0.94% (0.77-1.15)	13.9% (11.5-17.1)	2.2 (1.8-2.7)
Massachusetts	1.71% (1.36-2.13)	14.4% (11.6-18.0)	1.7 (1.4-2.1)
Michigan	1.00% (0.77-1.29)	17.1% (13.3-22.1)	2.6 (2.0-3.4)
Minnesota	0.64% (0.52-0.76)	18.4% (15.4-22.4)	2.1 (1.8-2.6)
Mississippi	0.74% (0.61-0.89)	31.6% (26.6-38.6)	3.2 (2.7-3.9)
Missouri	0.69% (0.55-0.87)	19.6% (15.8-24.7)	2.5 (2.0-3.2)
Montana	0.62% (0.50-0.75)	21.0% (17.5-25.5)	2.3 (1.9-2.8)

Table A.2: Posterior median and 95% intervals for IFR, cumulative incidence as of March 7, 2021, and undercount factor as of March 7, 2021 (continued).

State	IFR	Cumulative Incidence	Undercount
Nebraska	0.53% (0.43-0.64)	21.1% (17.7-25.7)	2.0 (1.7-2.5)
Nevada	0.73% (0.59-0.87)	22.5% (18.8-27.5)	2.4 (2.0-2.9)
New Hampshire	0.66% (0.54-0.79)	13.4% (11.3-16.2)	2.4 (2.0-2.9)
New Jersey	1.22% (0.97-1.53)	22.2% (17.7-28.0)	2.4 (1.9-3.1)
New Mexico	1.01% (0.83-1.21)	18.6% (15.6-22.7)	2.1 (1.8-2.5)
New York	1.12% (0.87-1.42)	18.6% (14.7-23.9)	2.1 (1.7-2.8)
North Carolina	0.58% (0.48-0.68)	19.4% (16.5-23.4)	2.4 (2.0-2.8)
North Dakota	0.86% (0.71-1.03)	22.4% (19.0-27.1)	1.7 (1.4-2.0)
Ohio	0.83% (0.68-1.03)	19.5% (15.9-23.7)	2.3 (1.9-2.9)
Oklahoma	0.47% (0.38-0.56)	25.1% (21.1-30.6)	2.3 (1.9-2.8)
Oregon	0.69% (0.55-0.85)	8.2% (6.7-10.0)	2.2 (1.8-2.8)
Pennsylvania	1.00% (0.82-1.19)	19.4% (16.4-23.7)	2.6 (2.2-3.2)
Rhode Island	1.41% (1.14-1.72)	17.4% (14.3-21.4)	1.4 (1.2-1.8)
South Carolina	0.76% (0.63-0.91)	23.0% (19.4-28.0)	2.3 (1.9-2.8)
South Dakota	0.68% (0.56-0.82)	31.2% (26.1-37.9)	2.5 (2.1-3.0)
Tennessee	0.79% (0.64-0.97)	21.8% (17.7-26.9)	1.9 (1.6-2.4)
Texas	0.70% (0.57-0.85)	22.7% (18.8-27.8)	2.5 (2.1-3.0)
Utah	0.22% (0.18-0.26)	28.0% (23.8-33.7)	2.5 (2.1-3.0)
Vermont	0.72% (0.57-0.92)	4.6% (3.8-5.6)	1.8 (1.5-2.2)
Virginia	1.53% (1.09-2.36)	8.3% (5.5-11.7)	1.2 (0.8-1.7)
Washington	0.51% (0.41-0.64)	12.9% (10.3-16.1)	2.9 (2.3-3.6)
West Virginia	0.86% (0.71-1.01)	15.4% (13.2-18.5)	2.0 (1.8-2.5)
Wisconsin	0.57% (0.46-0.69)	21.5% (17.9-26.6)	2.0 (1.7-2.5)
Wyoming	0.58% (0.46-0.71)	21.0% (17.3-25.9)	2.2 (1.8-2.7)
District of Columbia	1.19% (0.94-1.49)	12.2% (9.8-15.3)	2.1 (1.7-2.7)

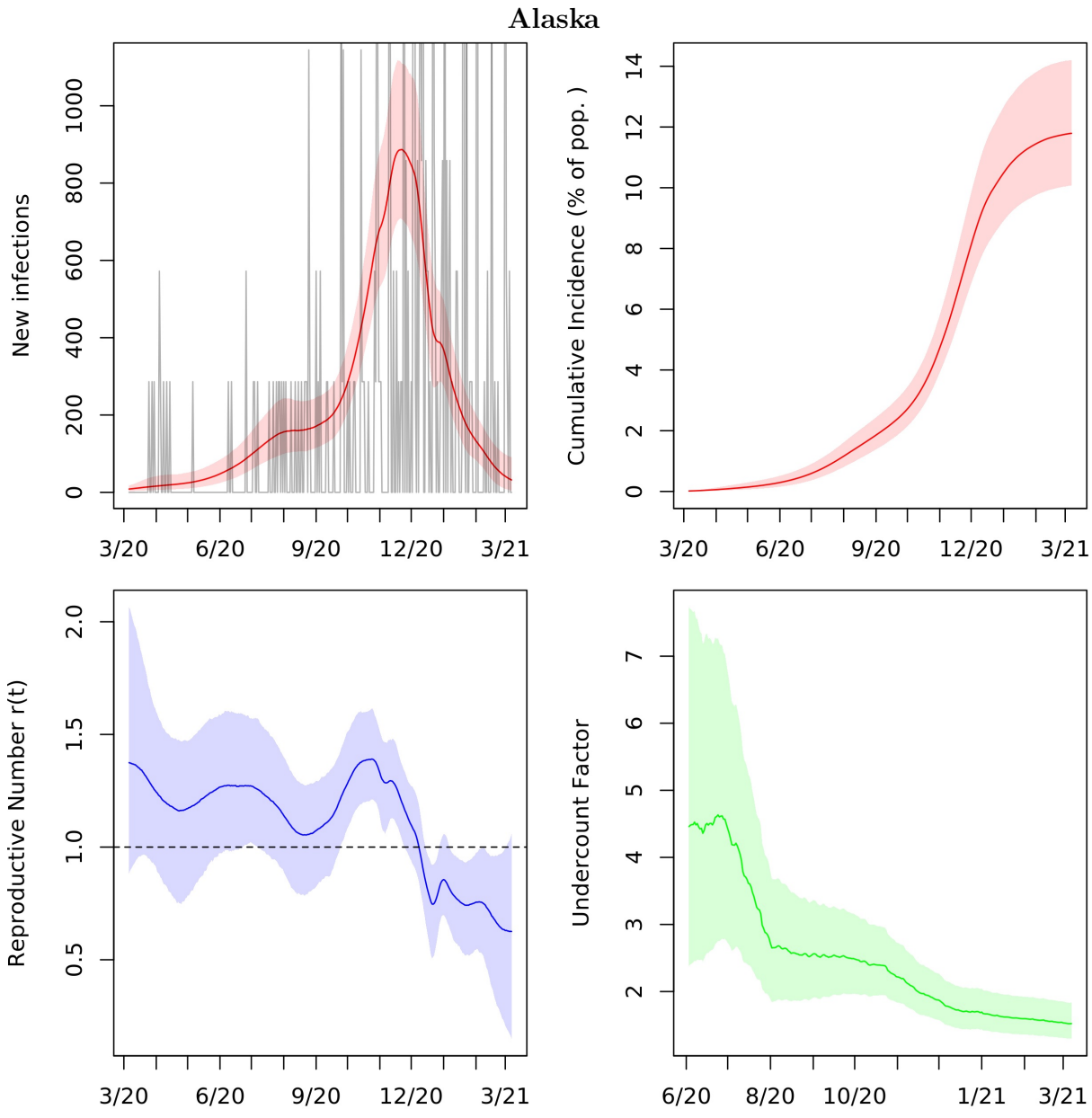


Figure A.1: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

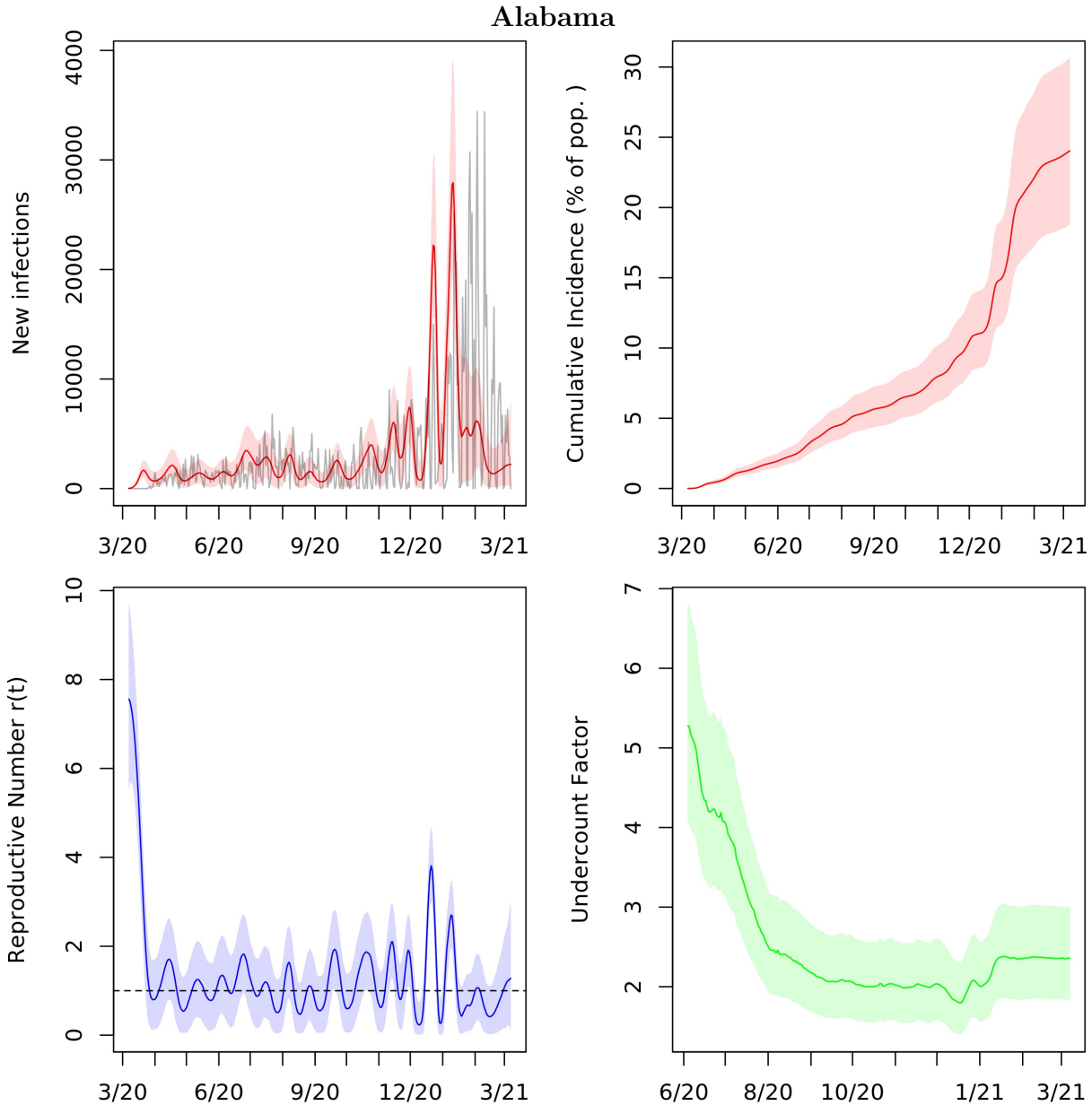


Figure A.2: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

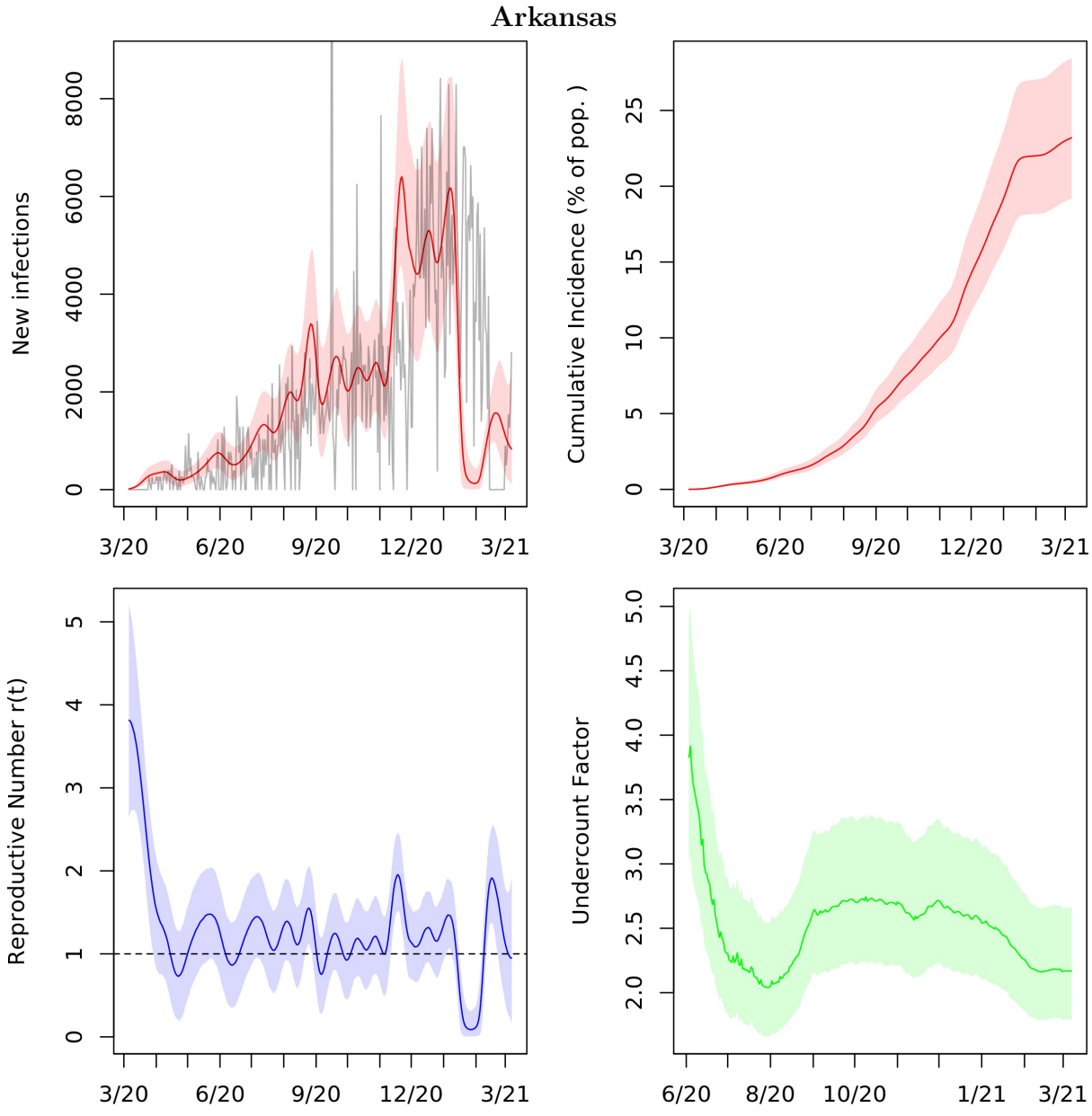


Figure A.3: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

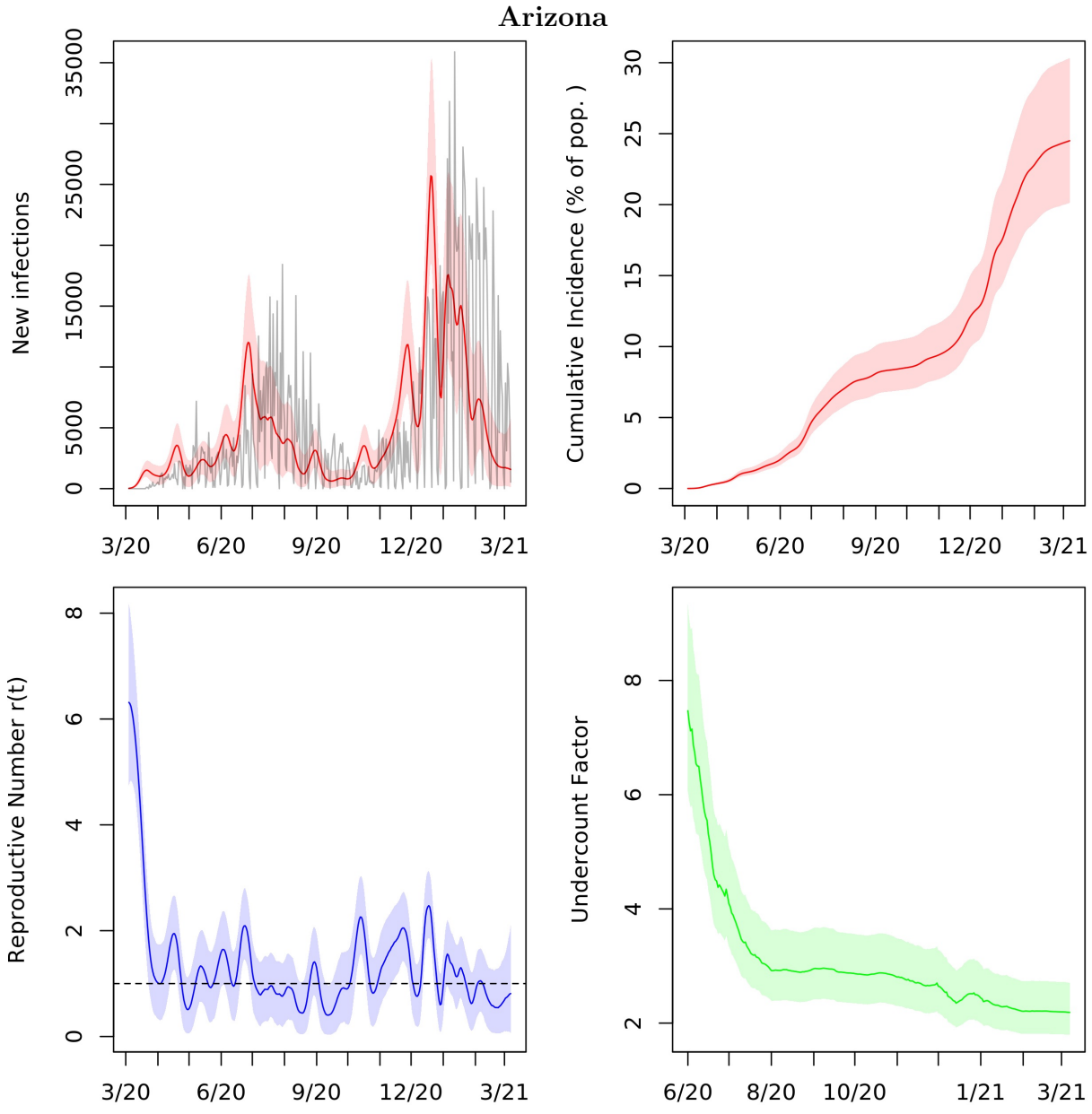


Figure A.4: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

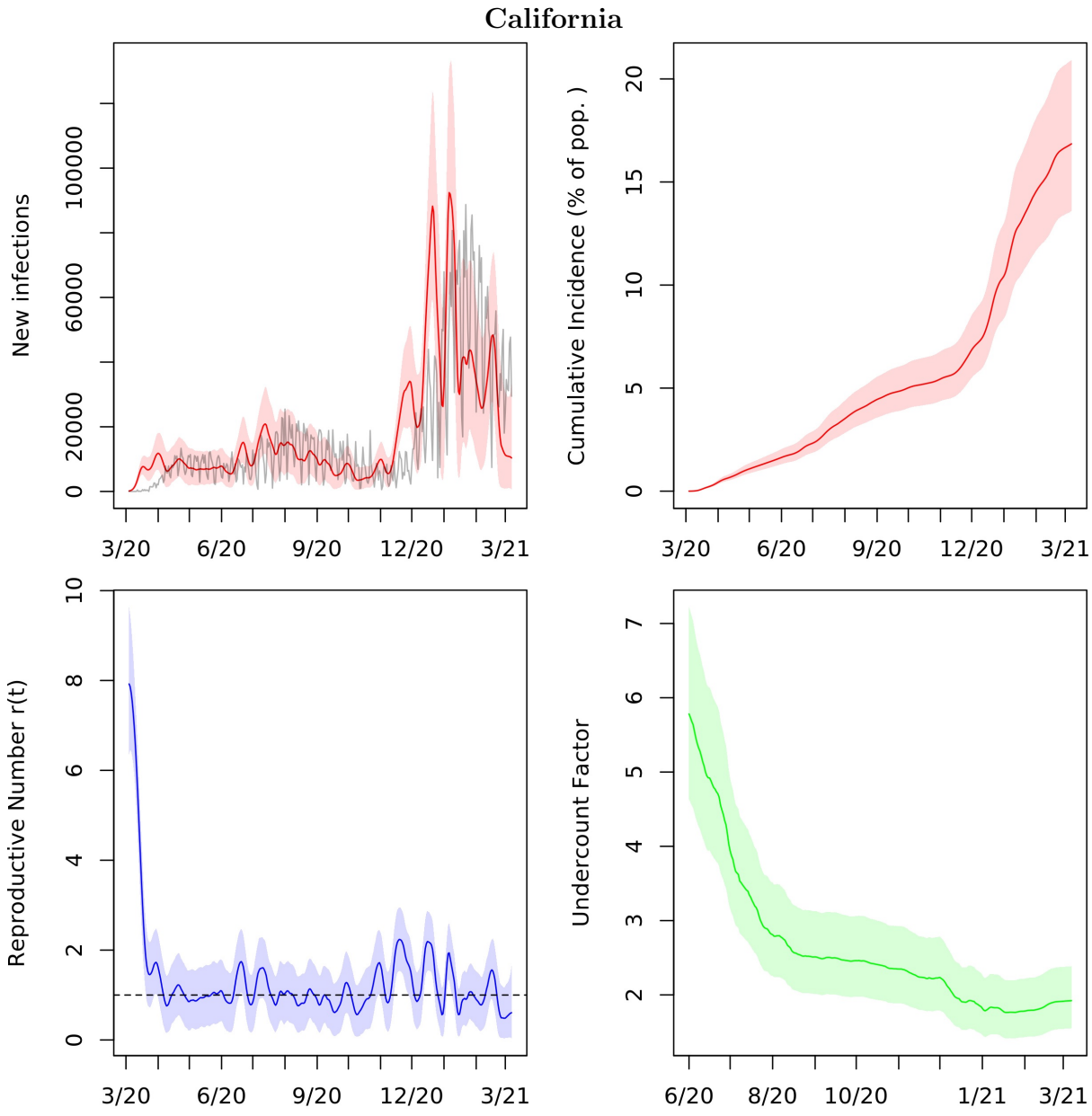


Figure A.5: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

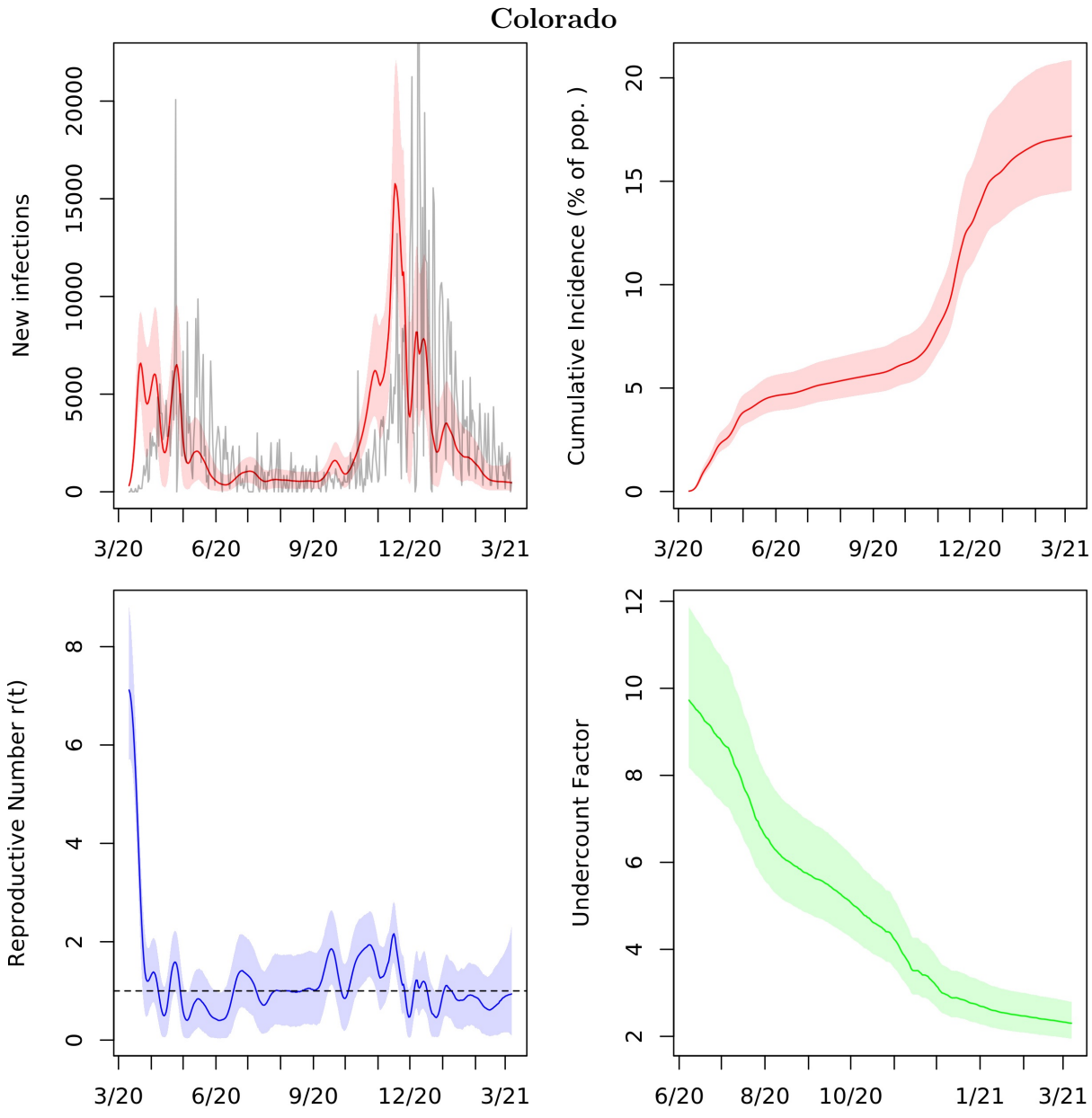


Figure A.6: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

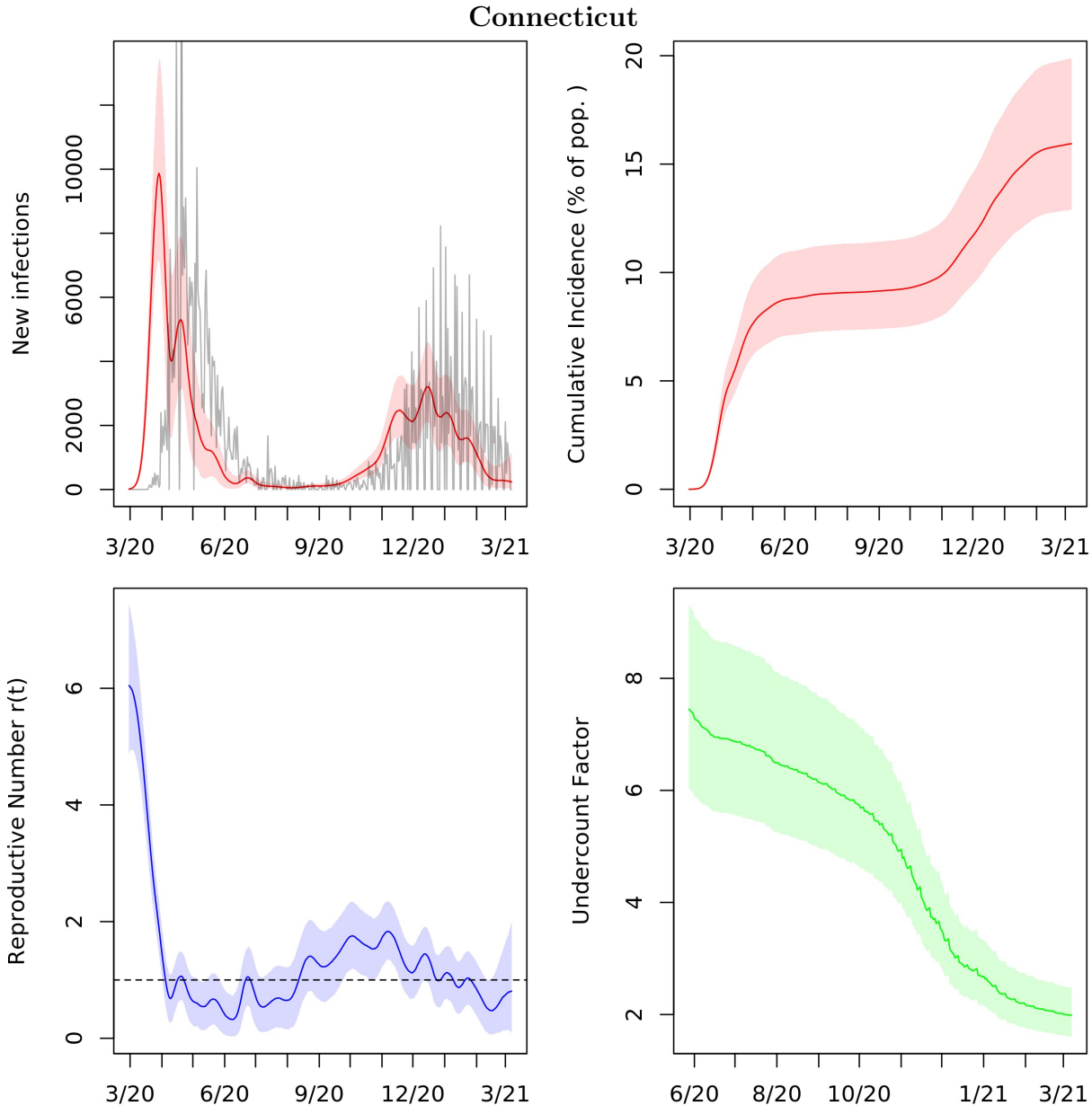


Figure A.7: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

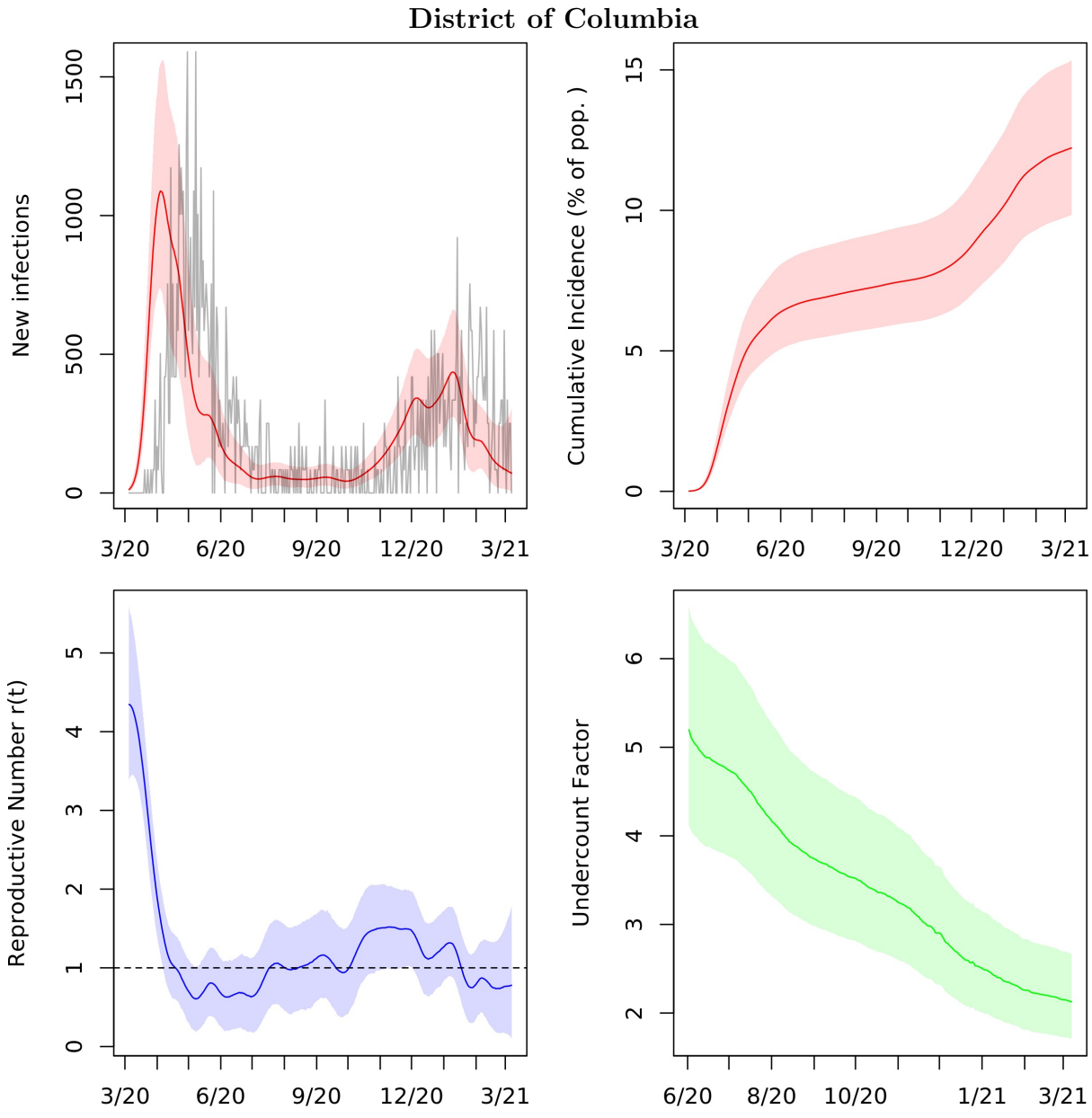


Figure A.8: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

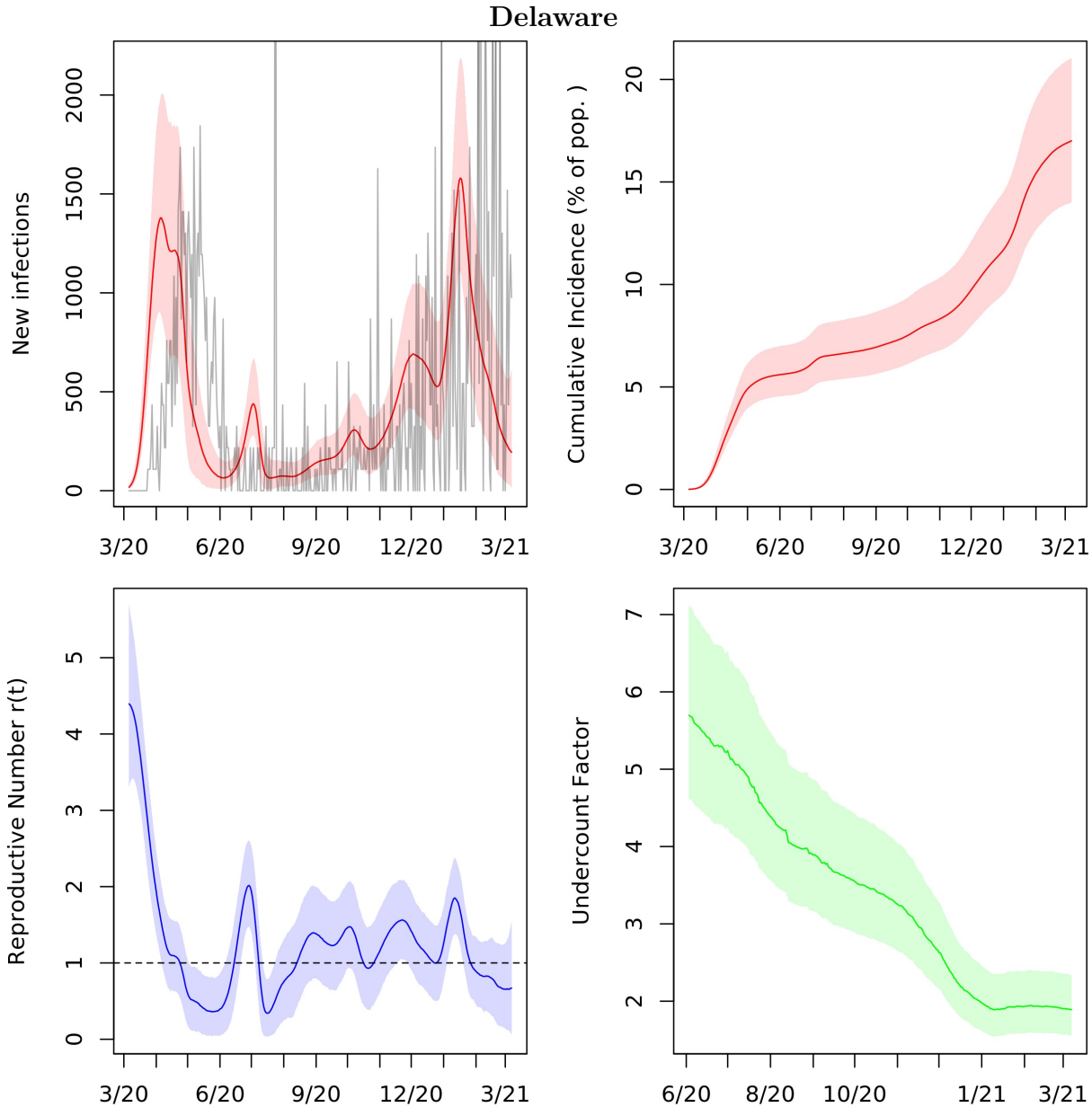


Figure A.9: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

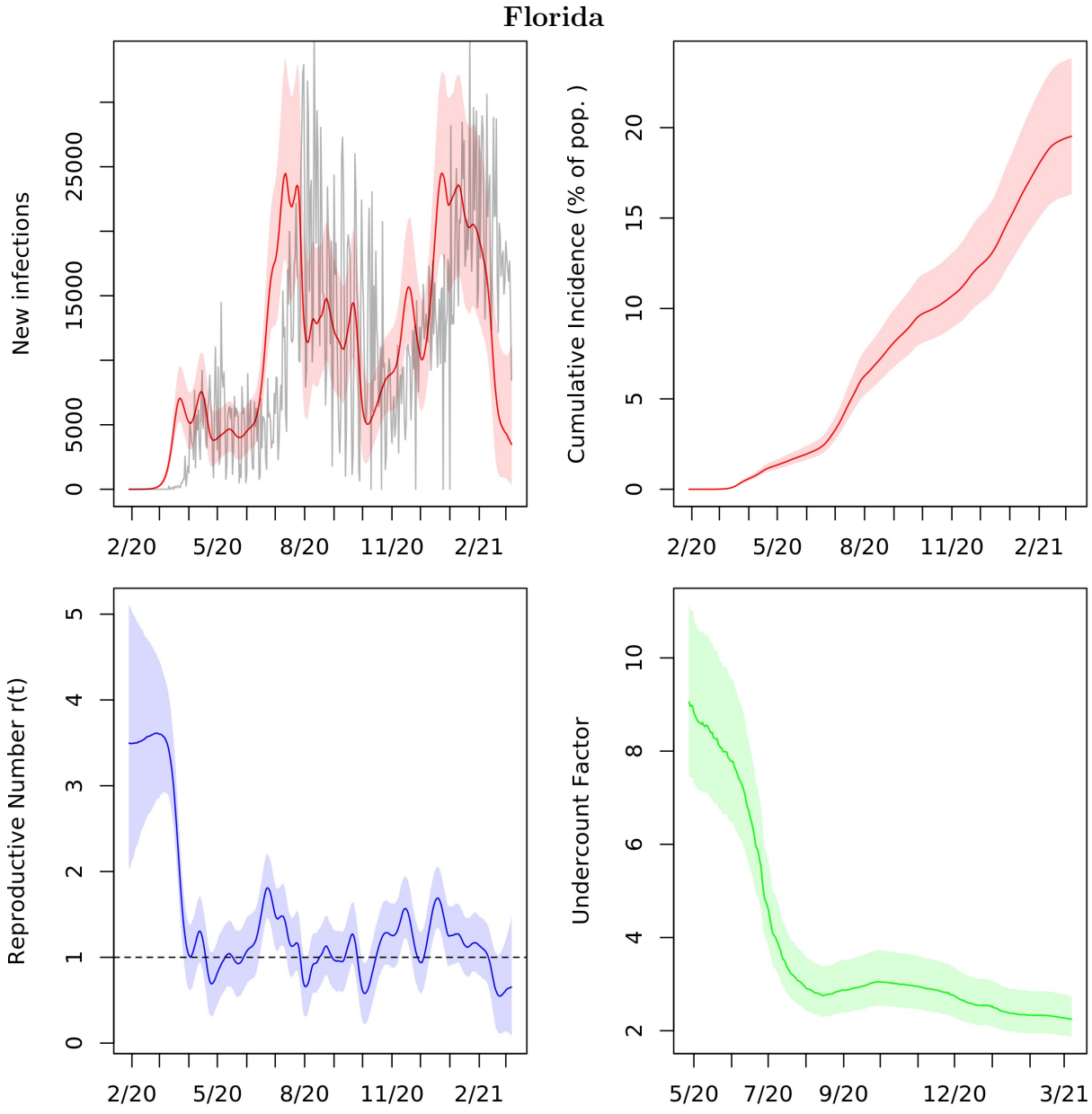


Figure A.10: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

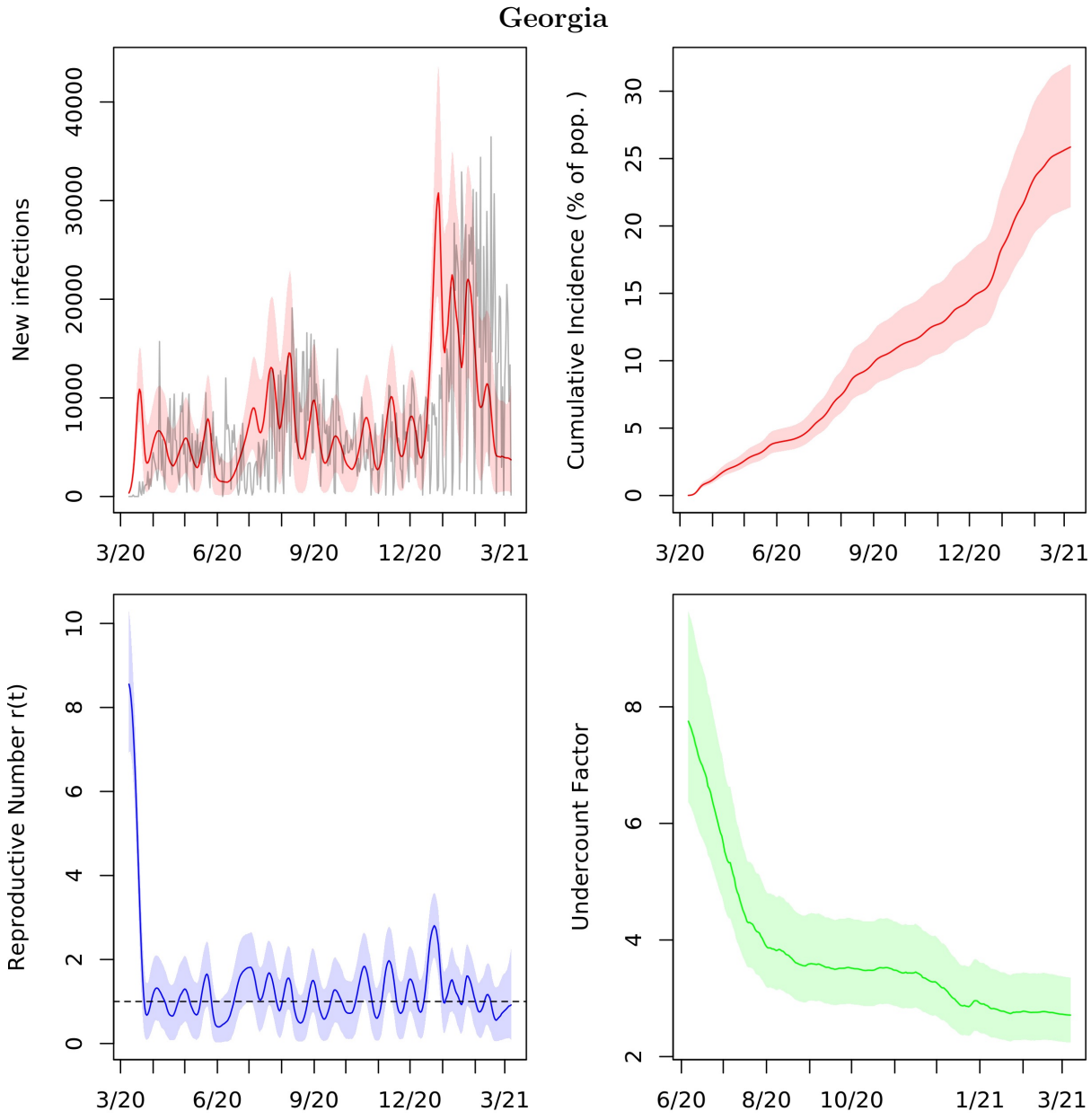


Figure A.11: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

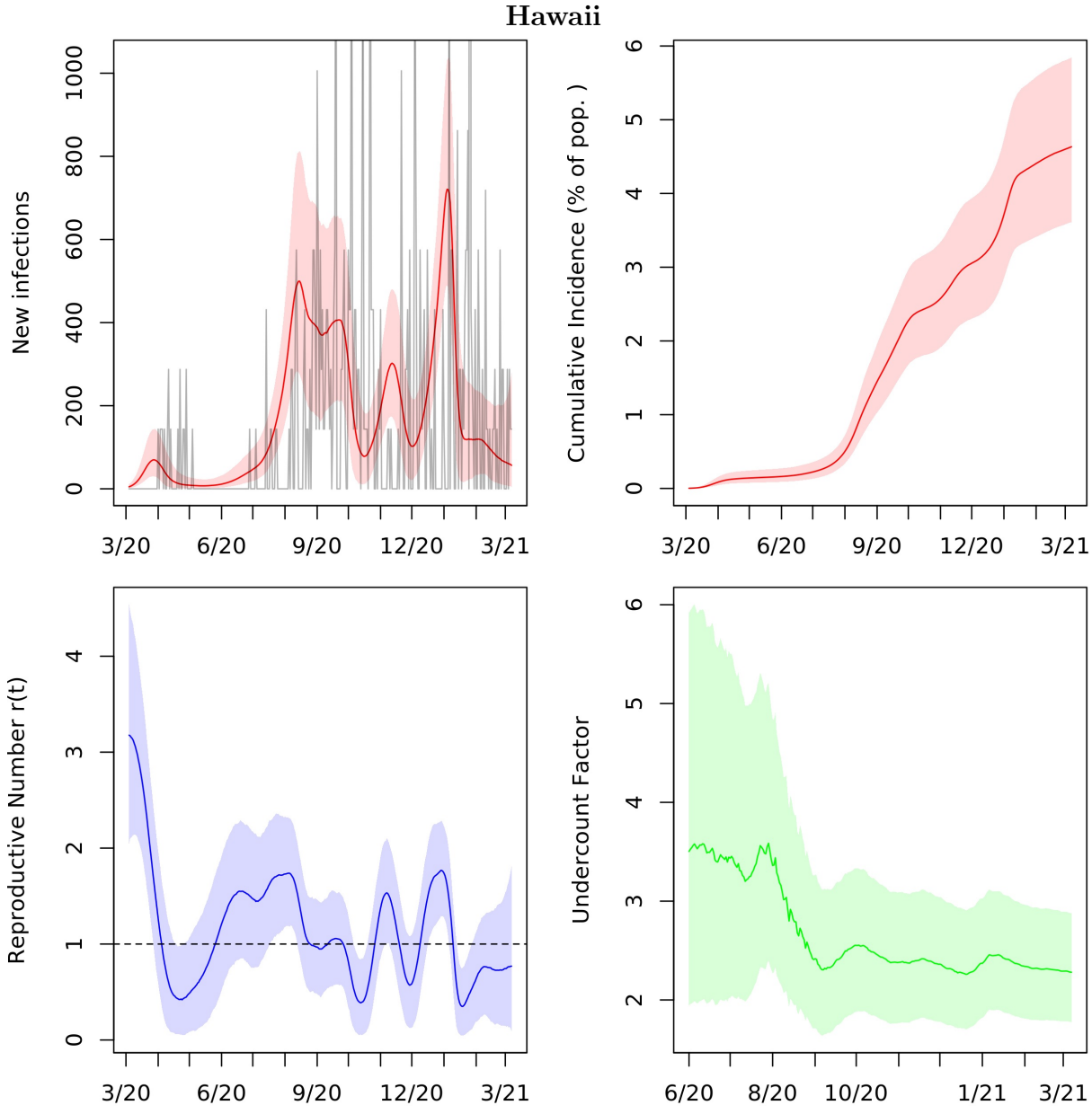


Figure A.12: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

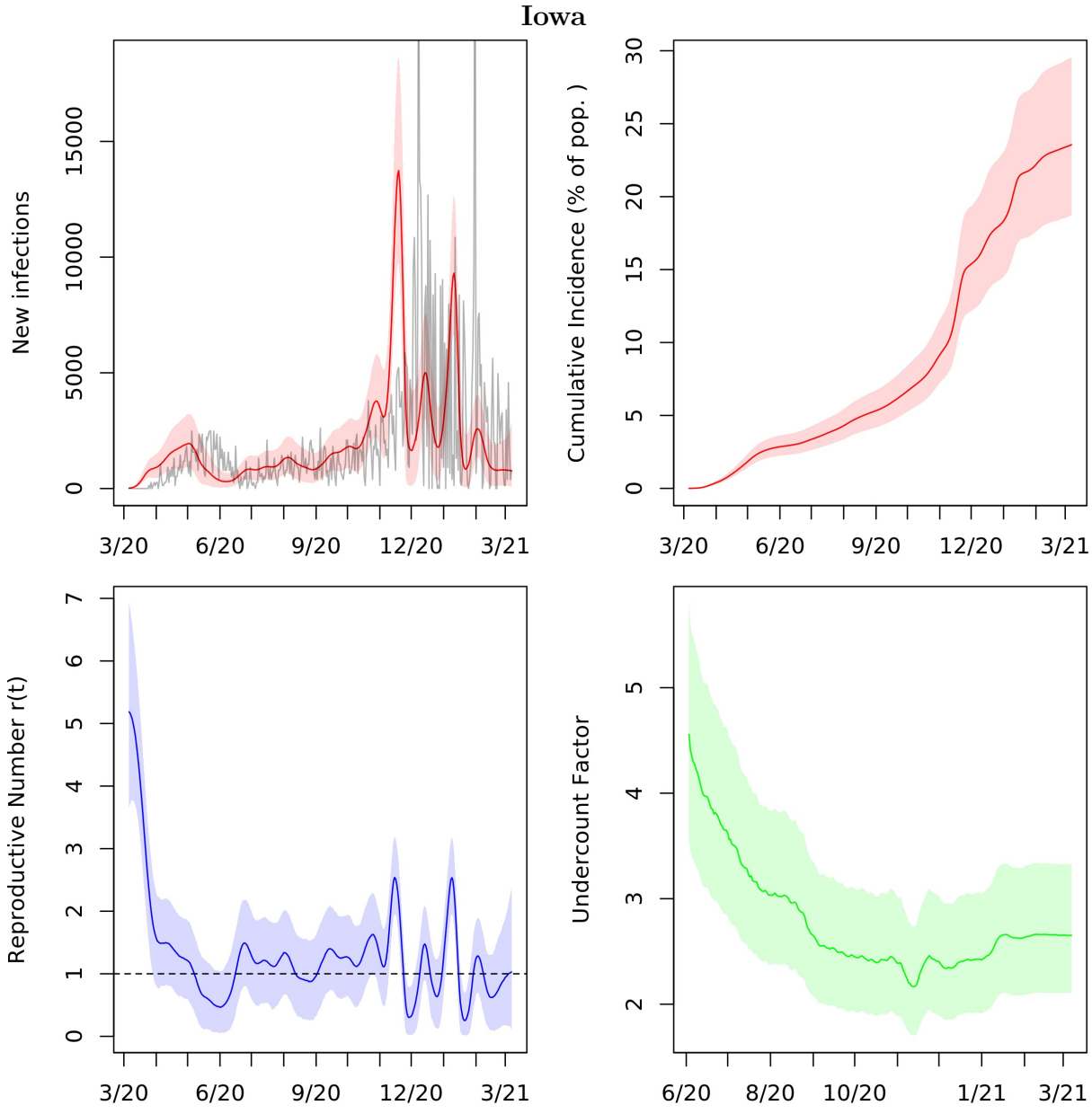


Figure A.13: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

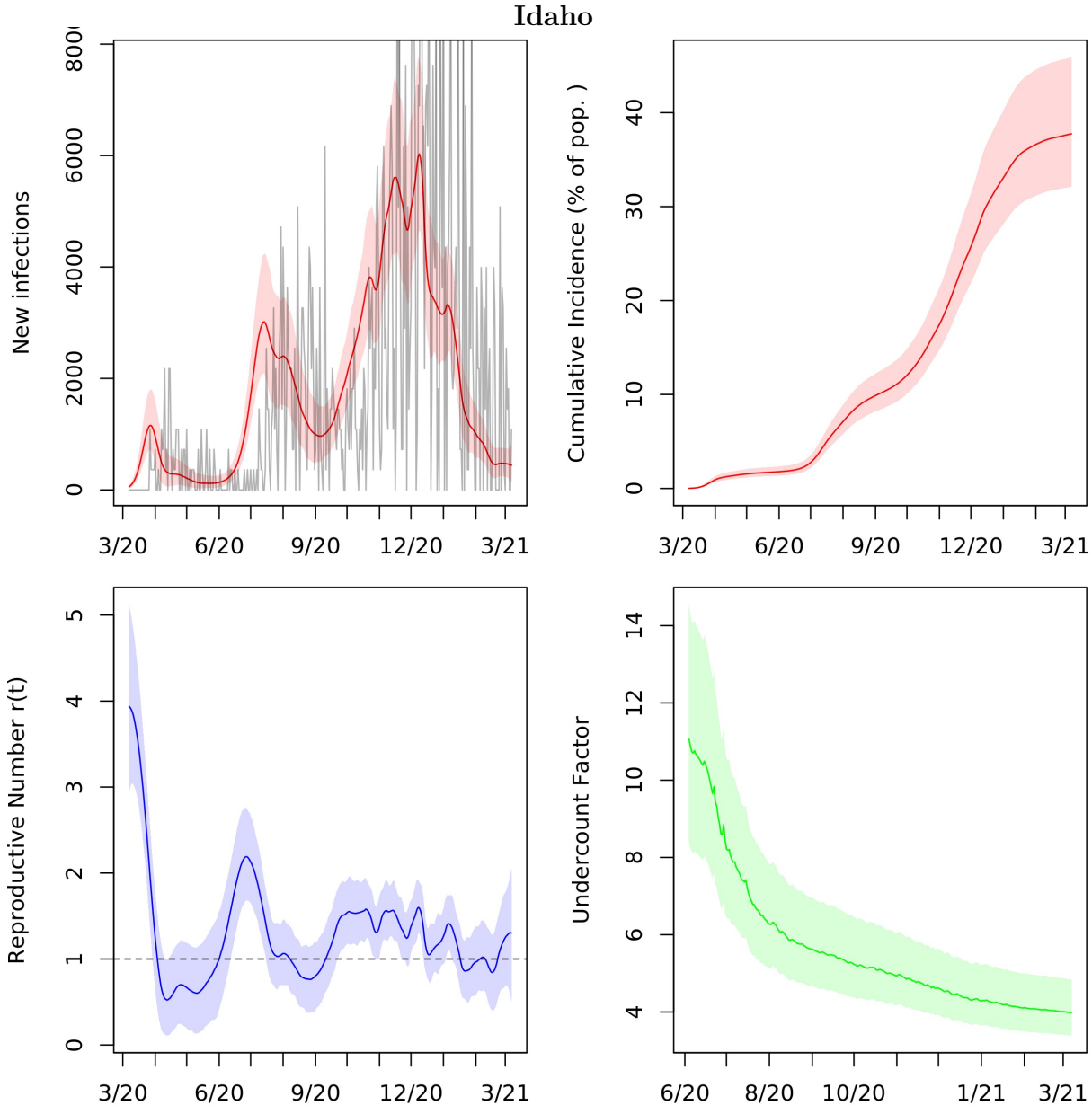


Figure A.14: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

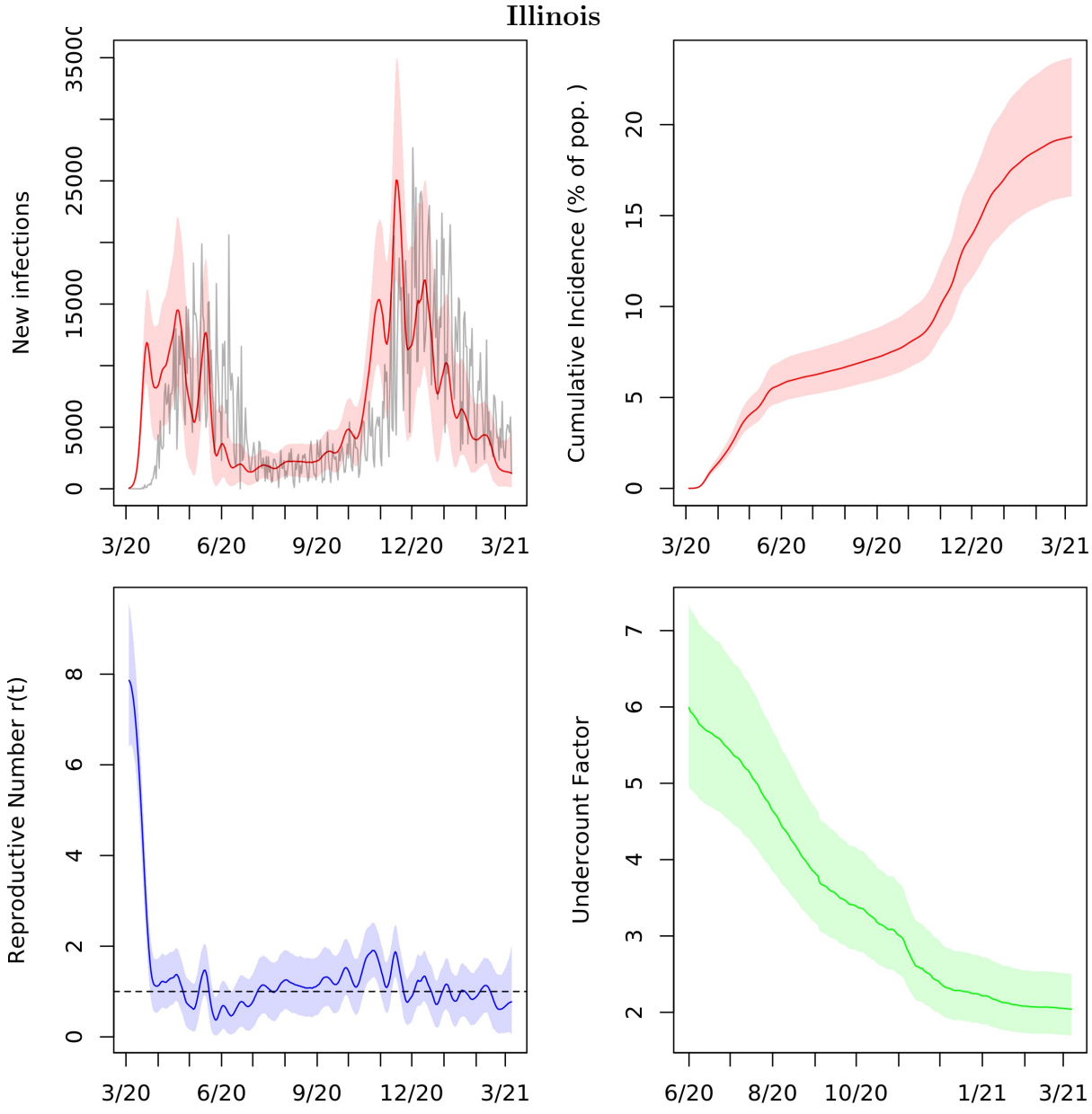


Figure A.15: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

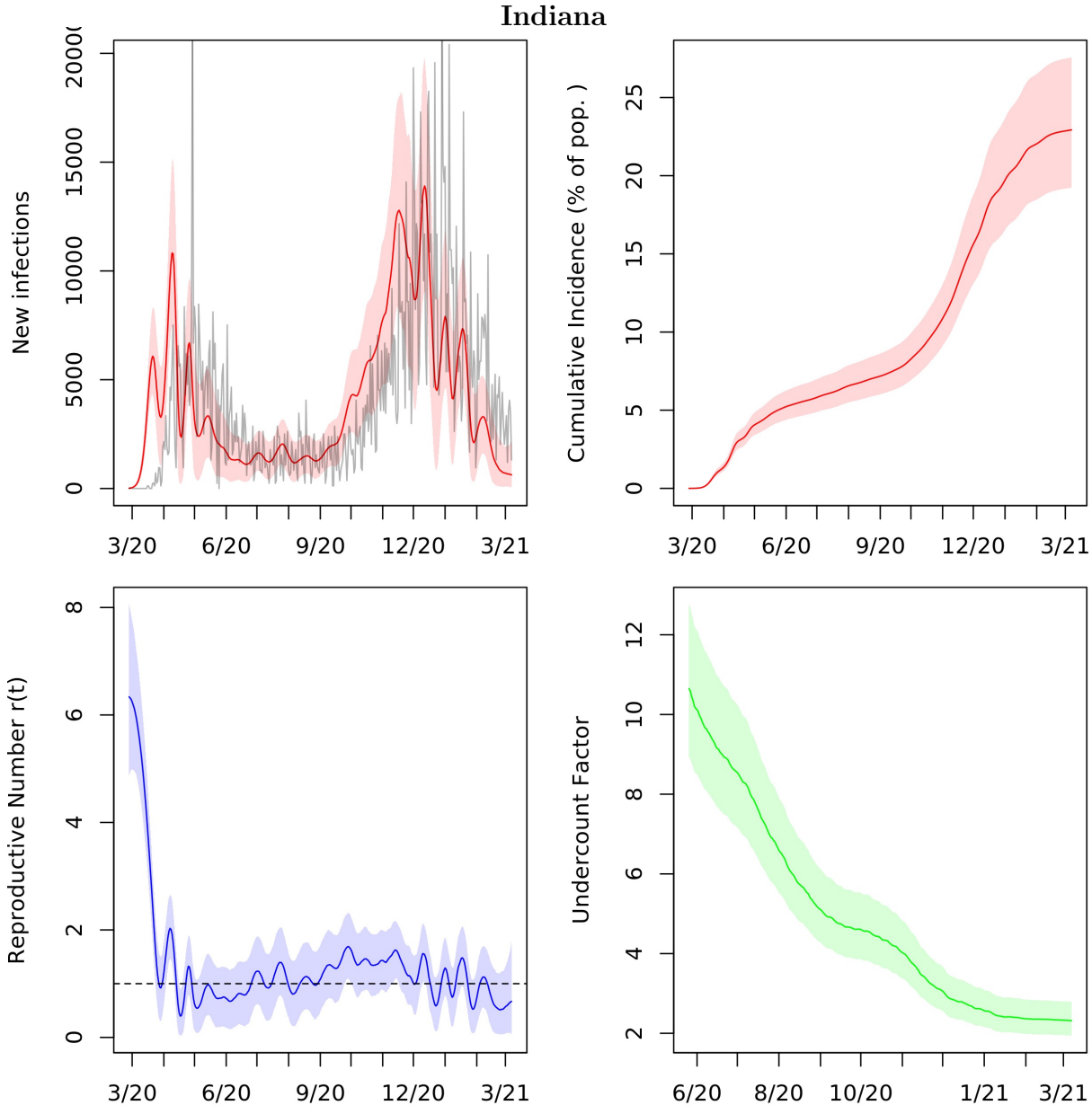


Figure A.16: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

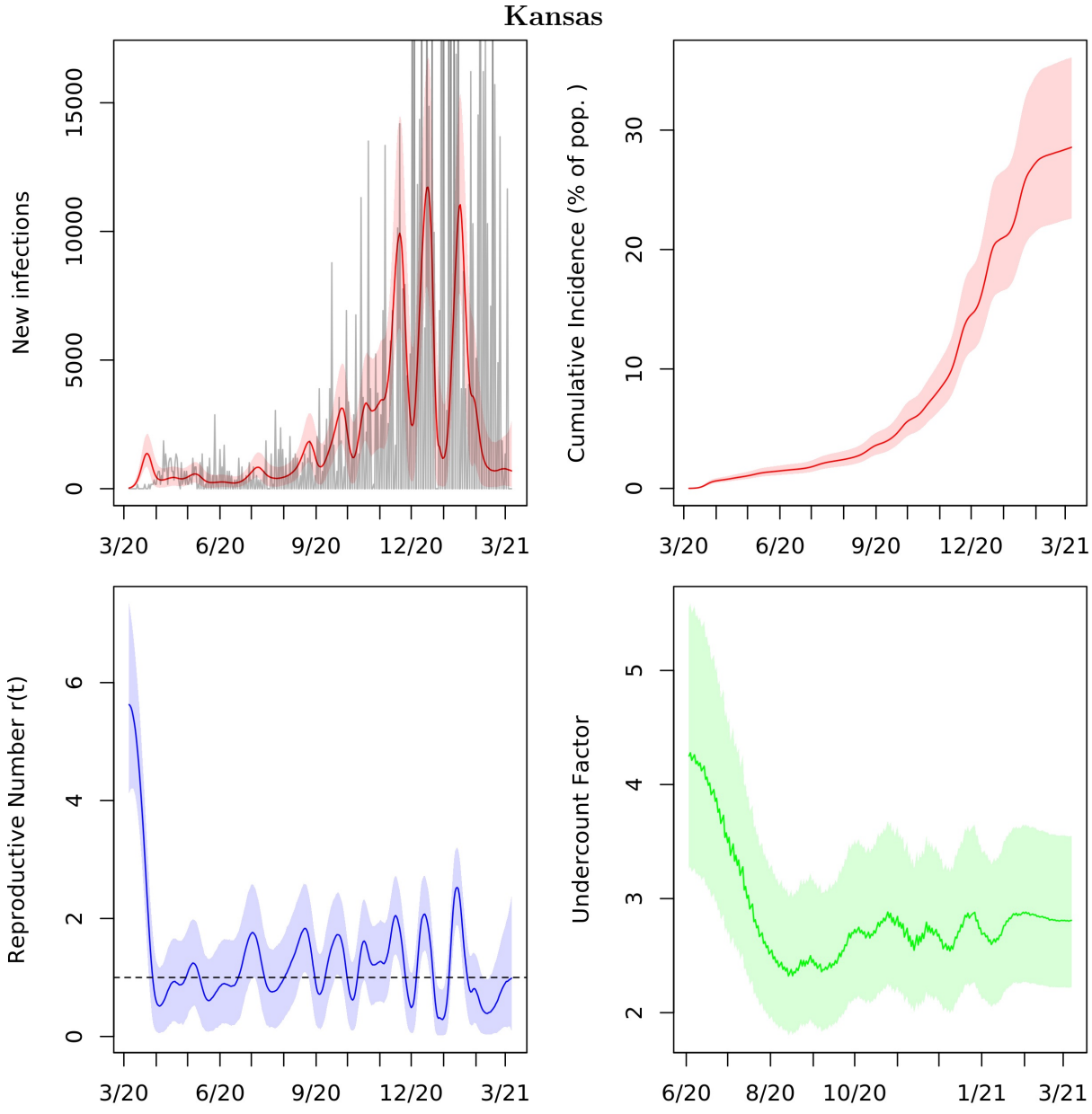


Figure A.17: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

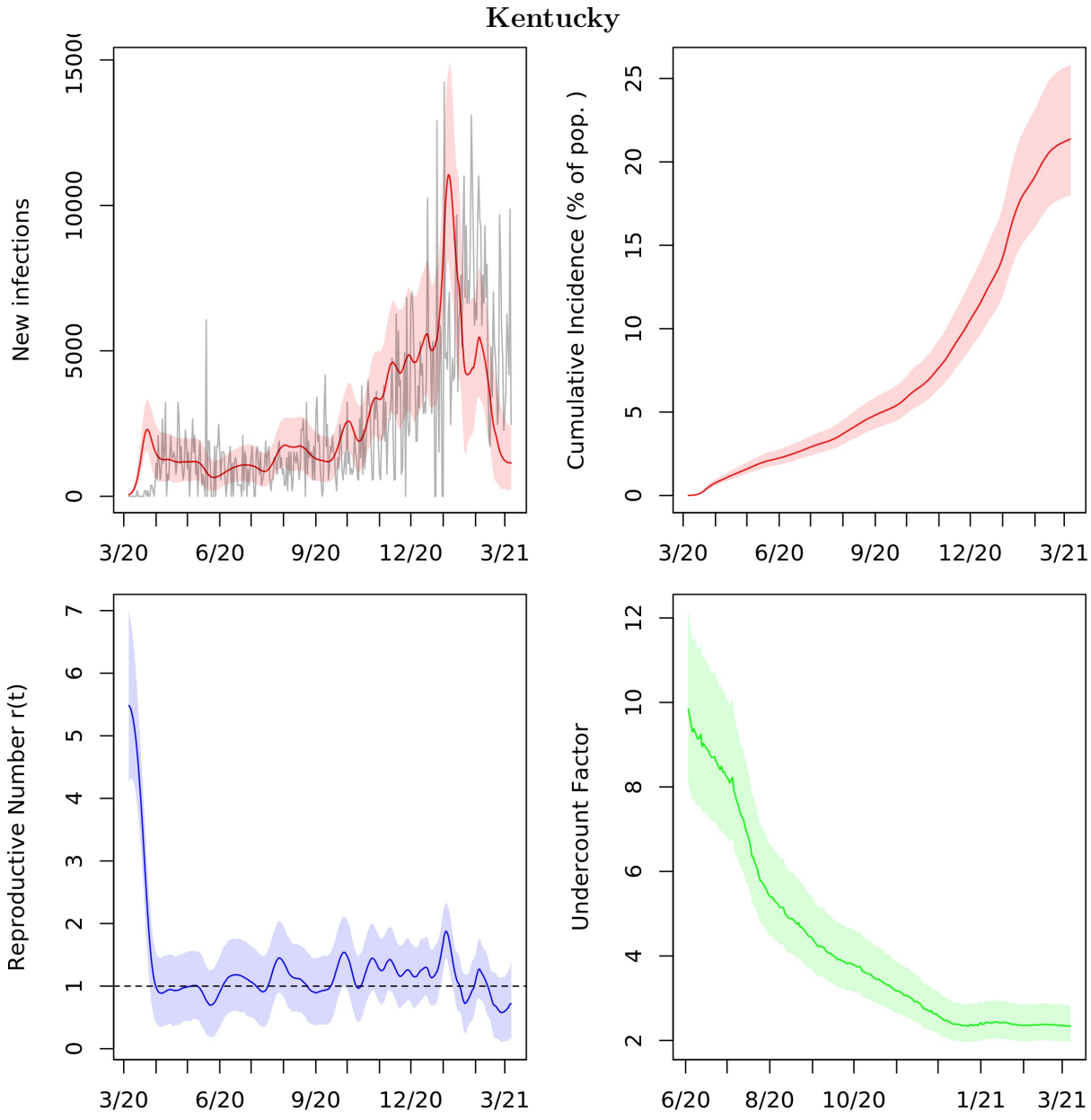


Figure A.18: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

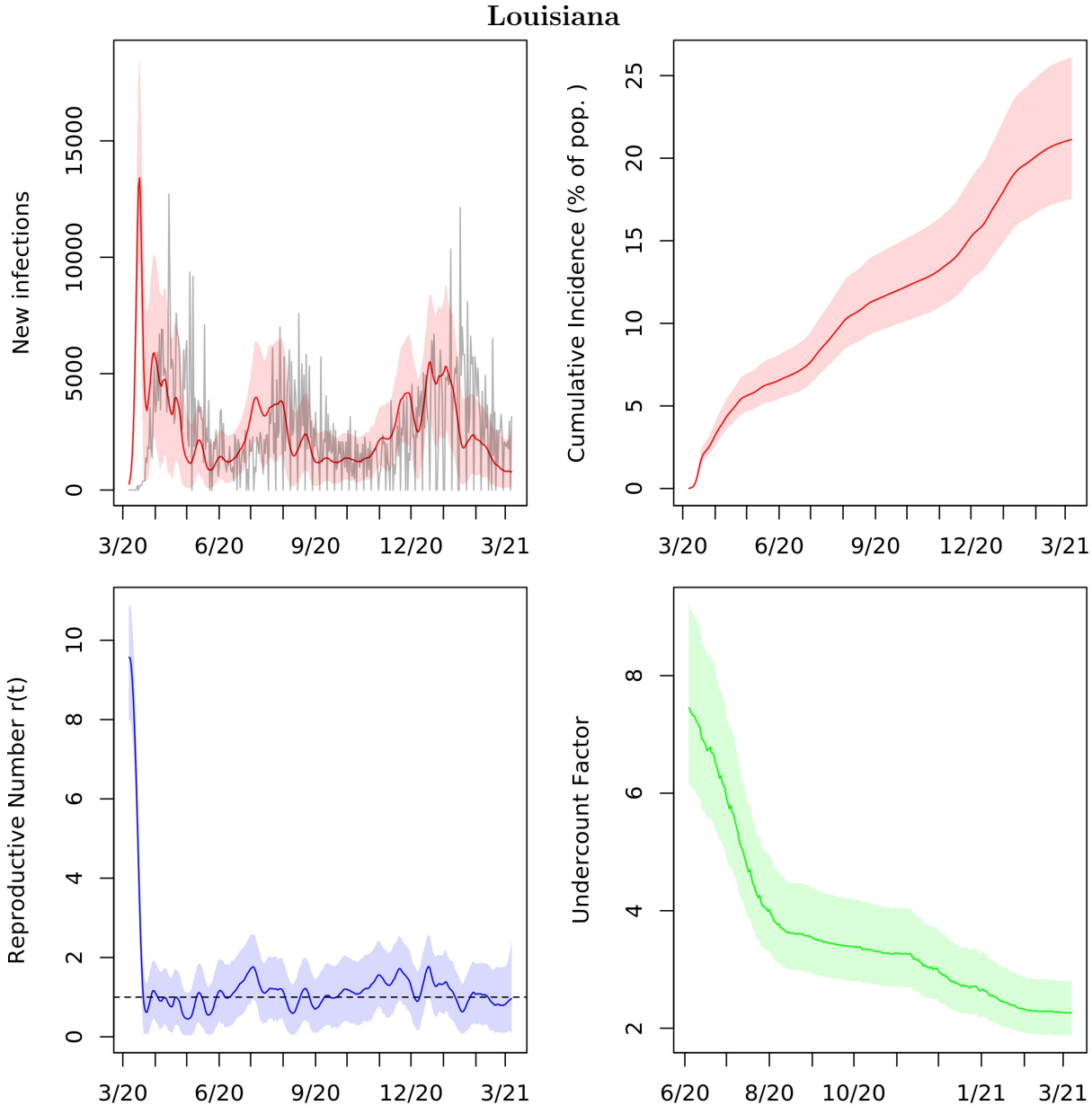


Figure A.19: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

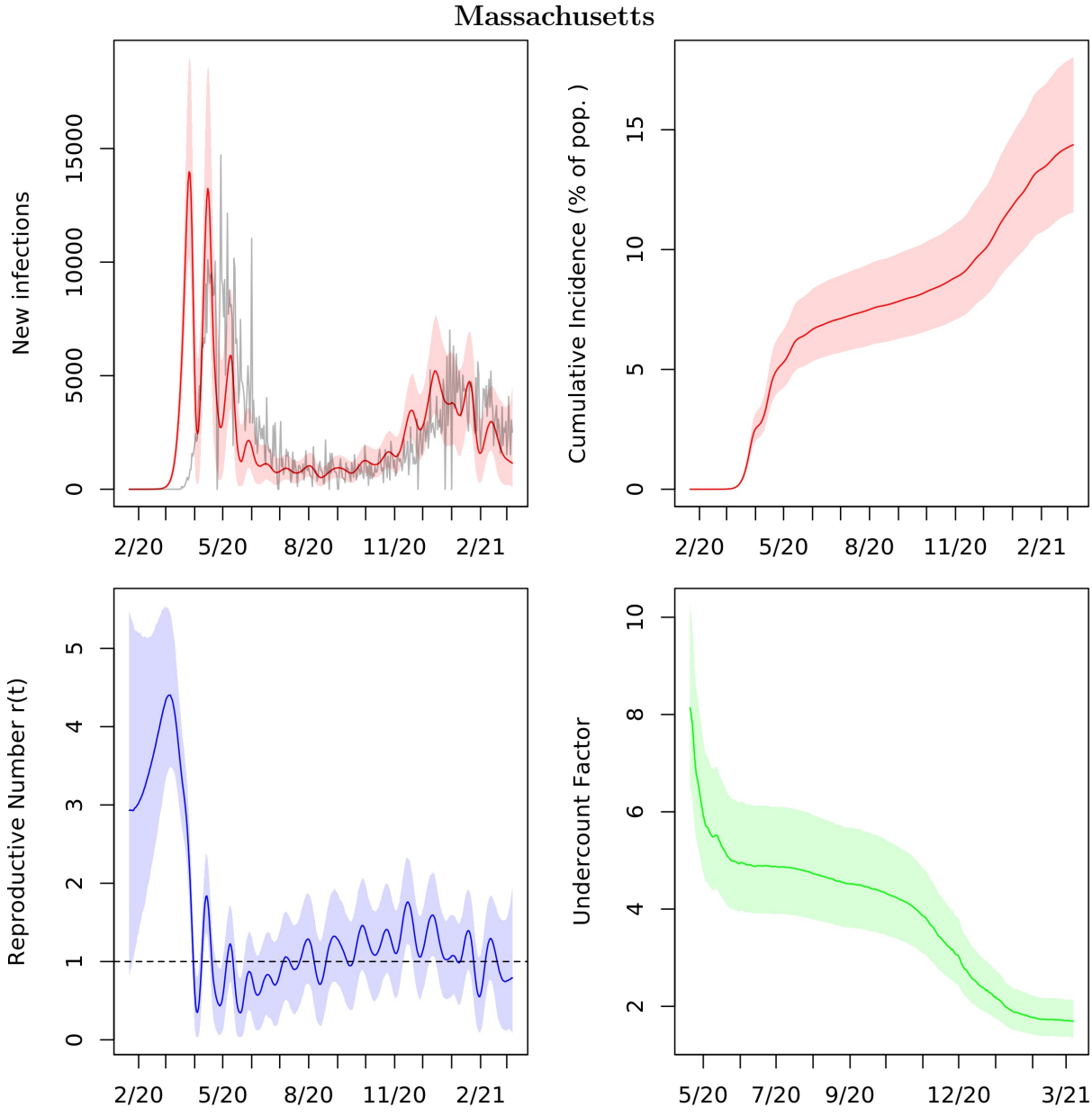


Figure A.20: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

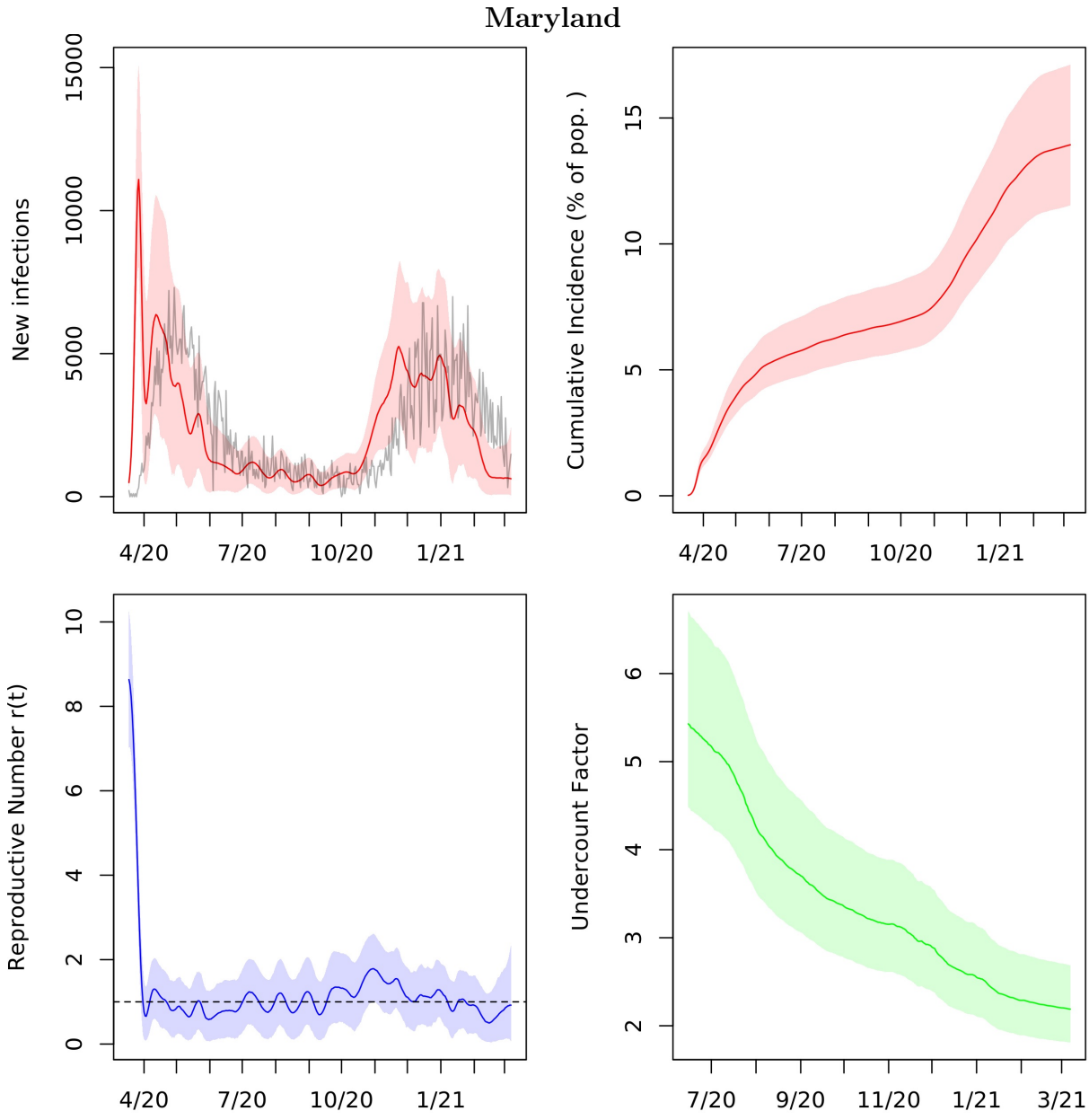


Figure A.21: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

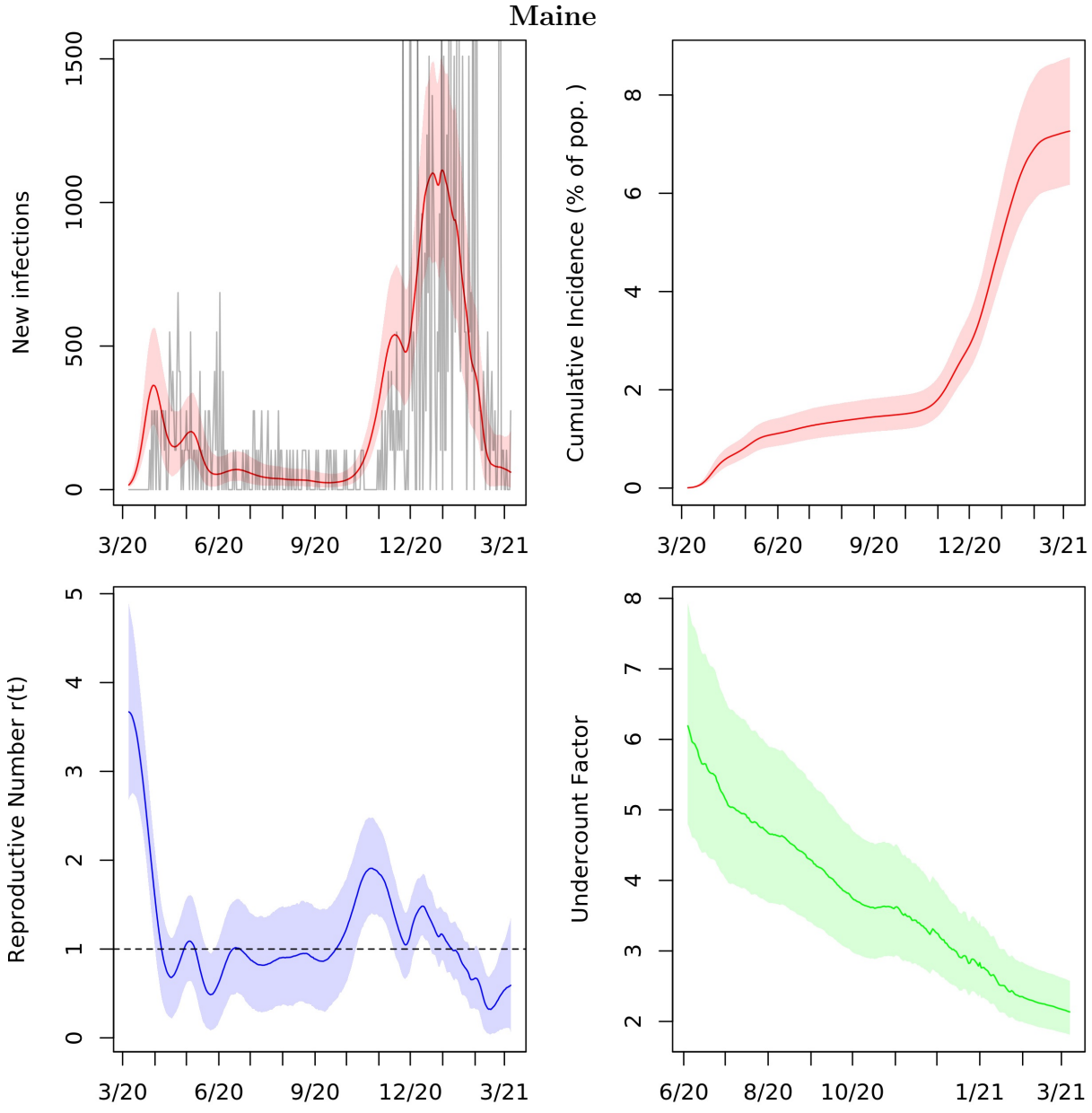


Figure A.22: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

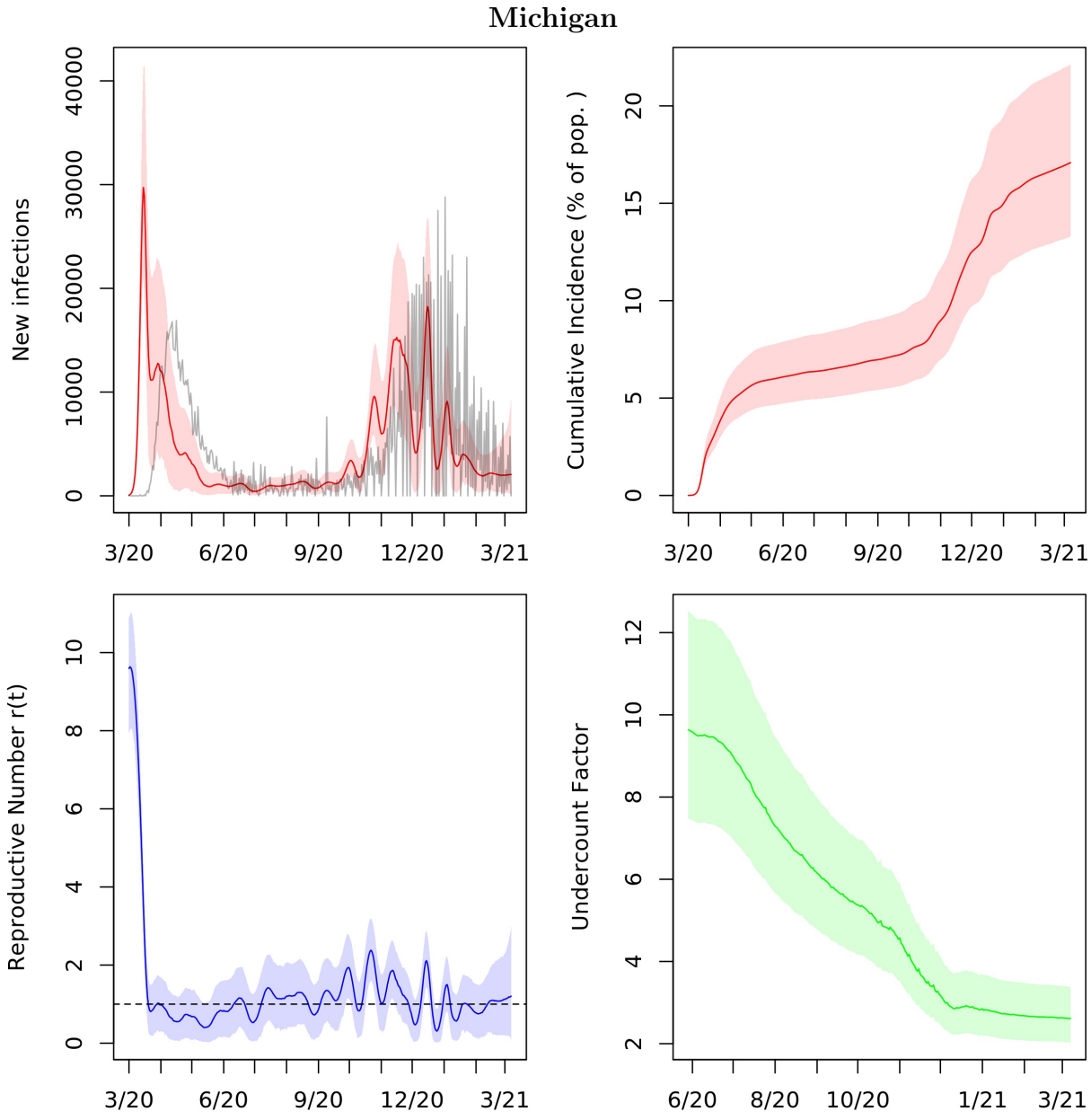


Figure A.23: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

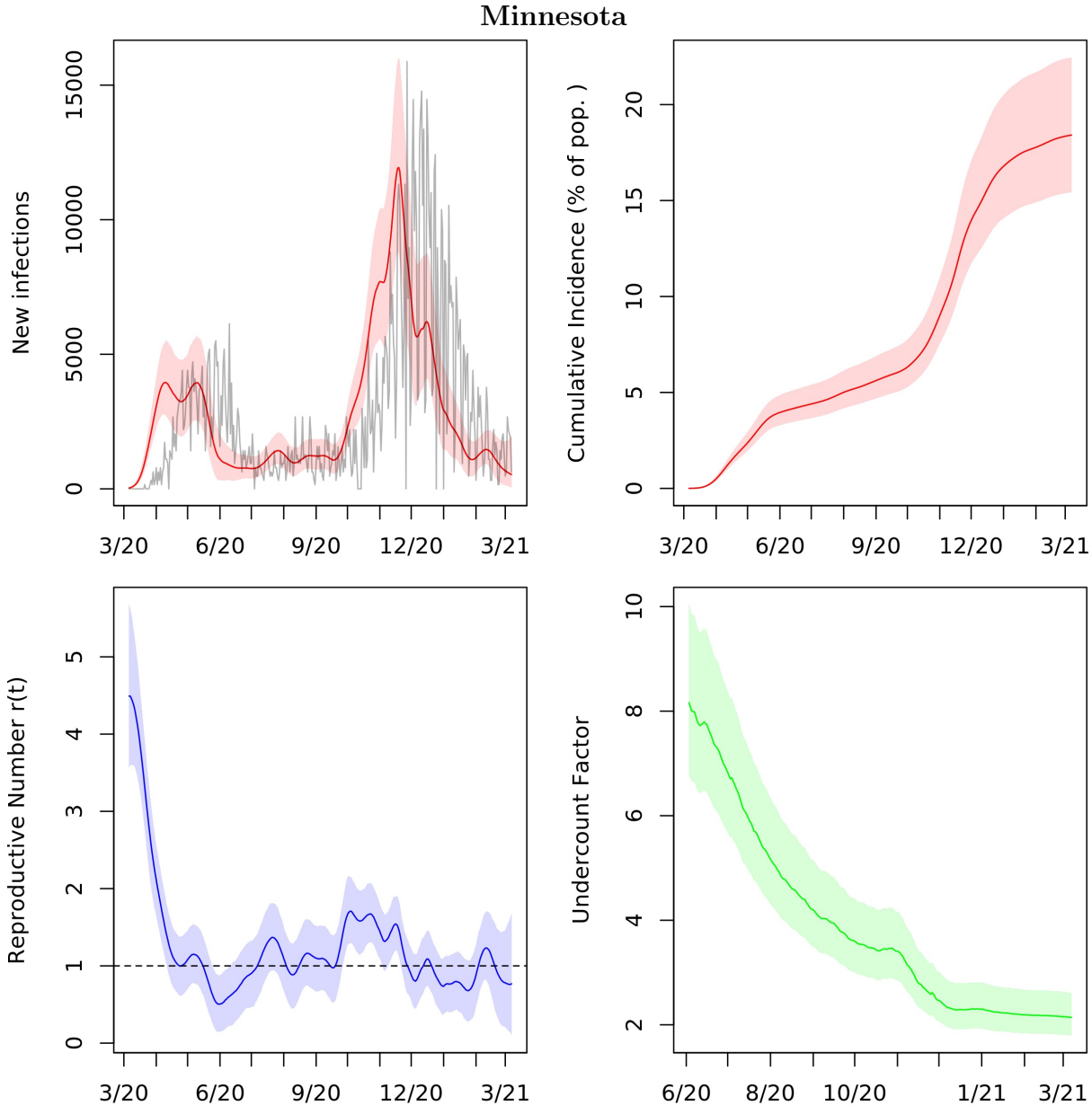


Figure A.24: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

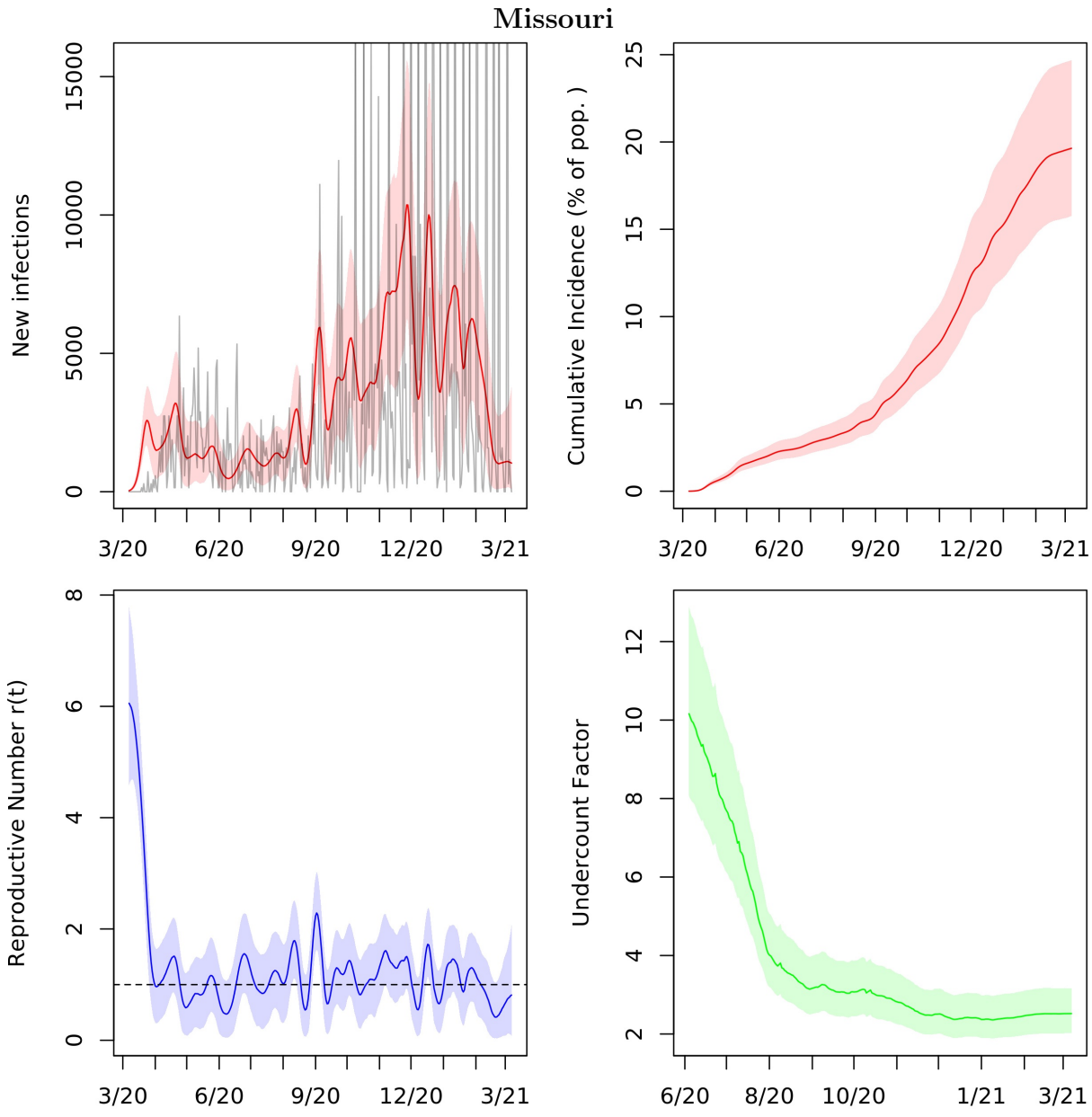


Figure A.25: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

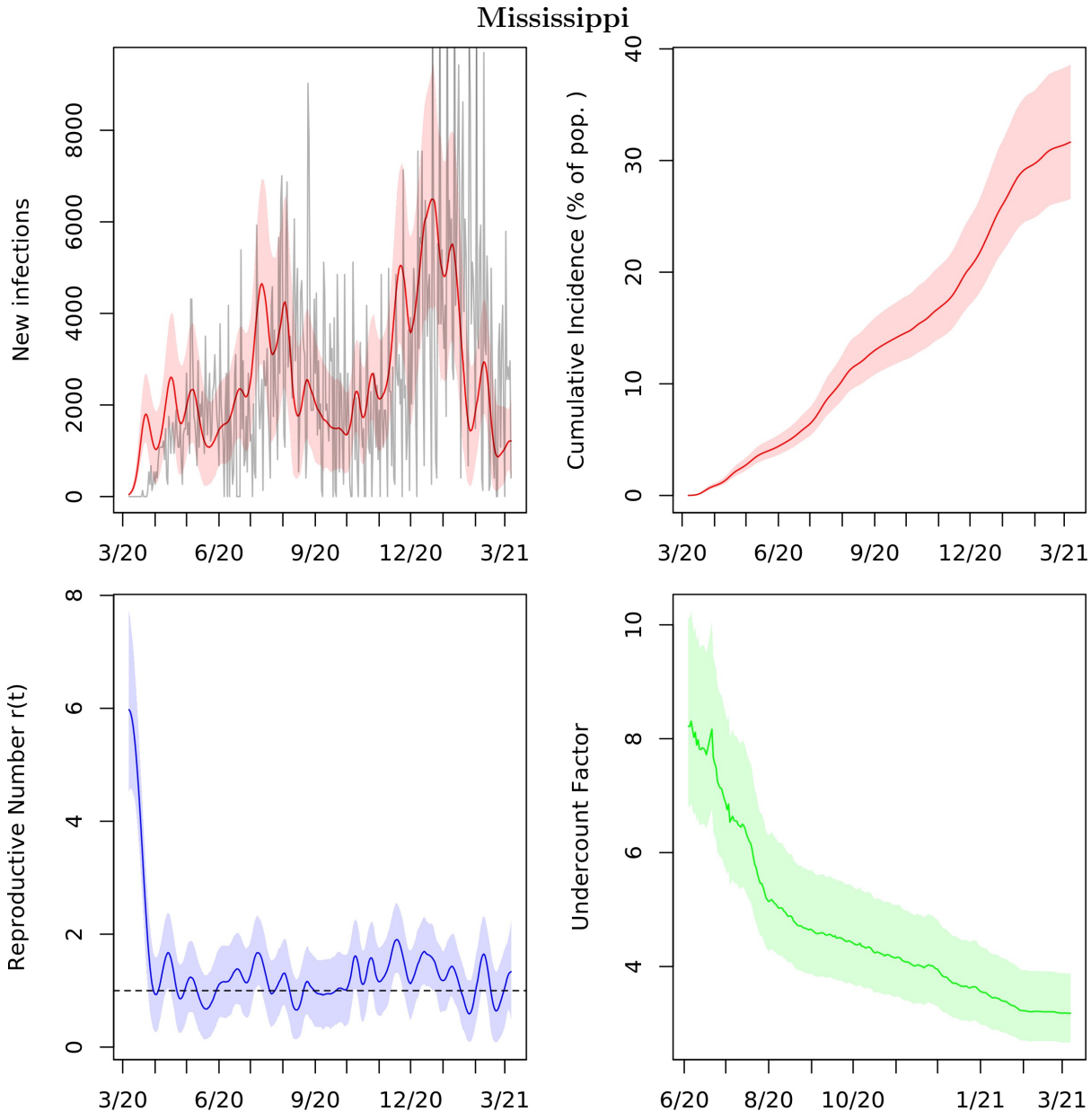


Figure A.26: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

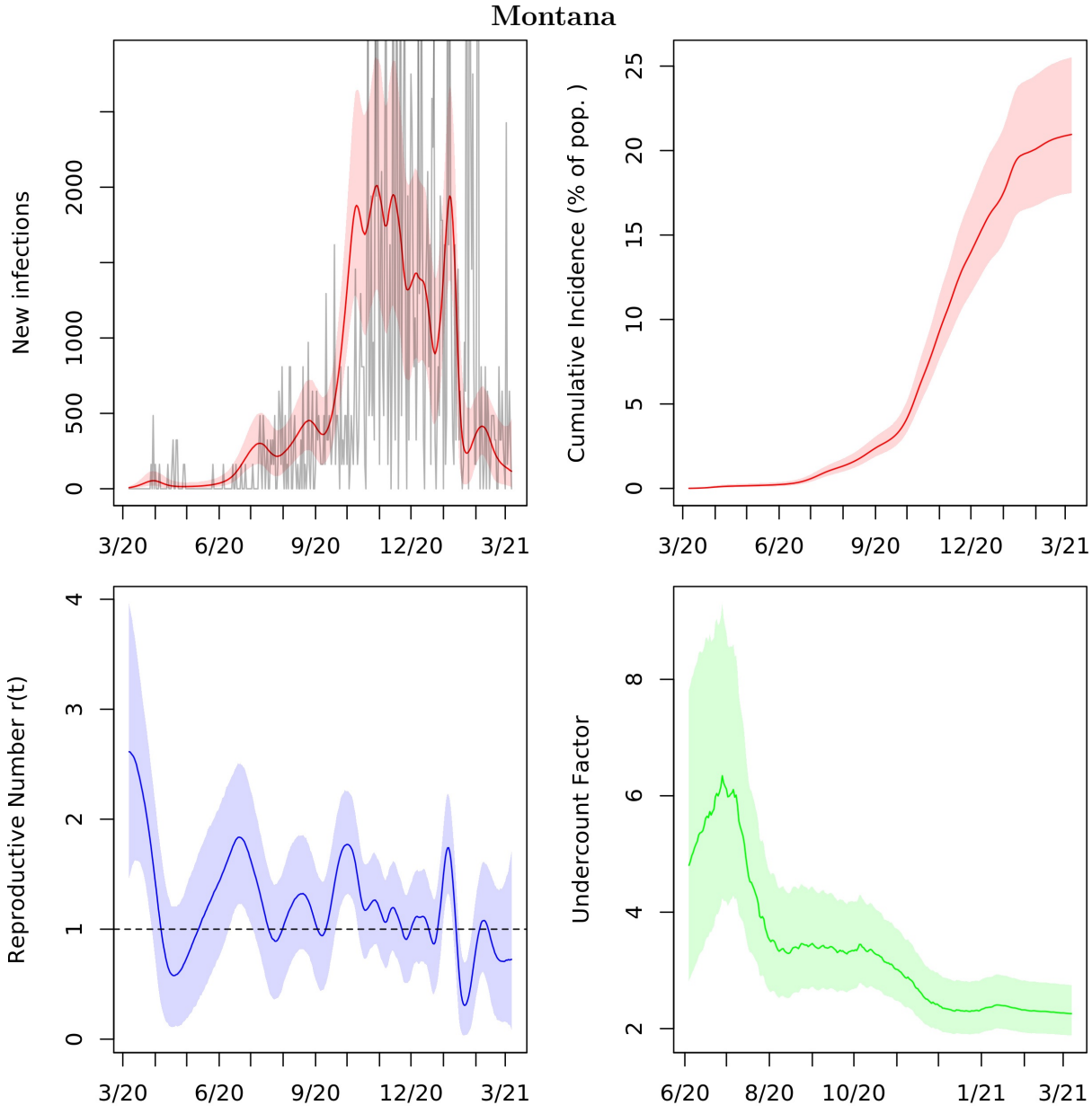


Figure A.27: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

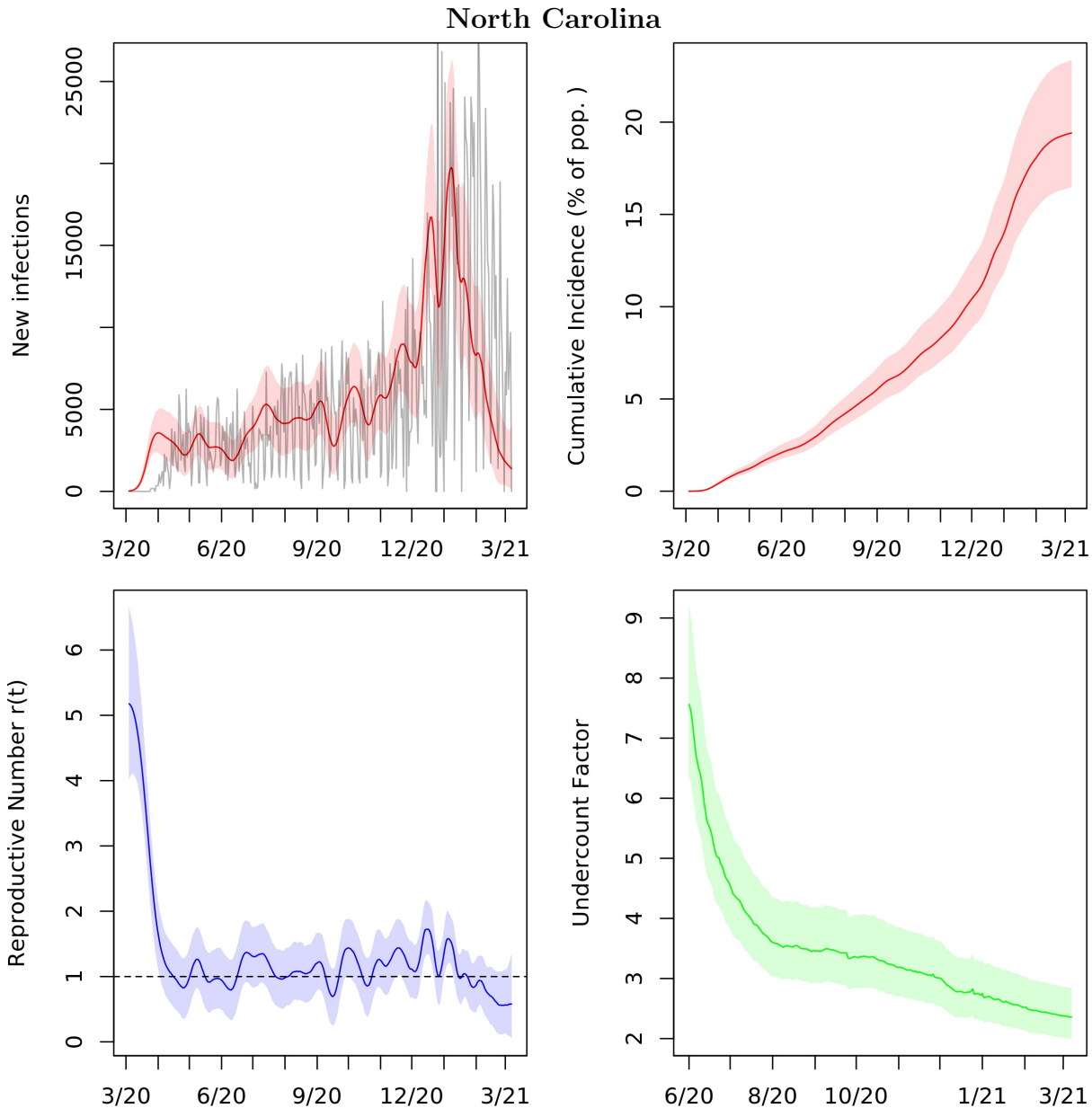


Figure A.28: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

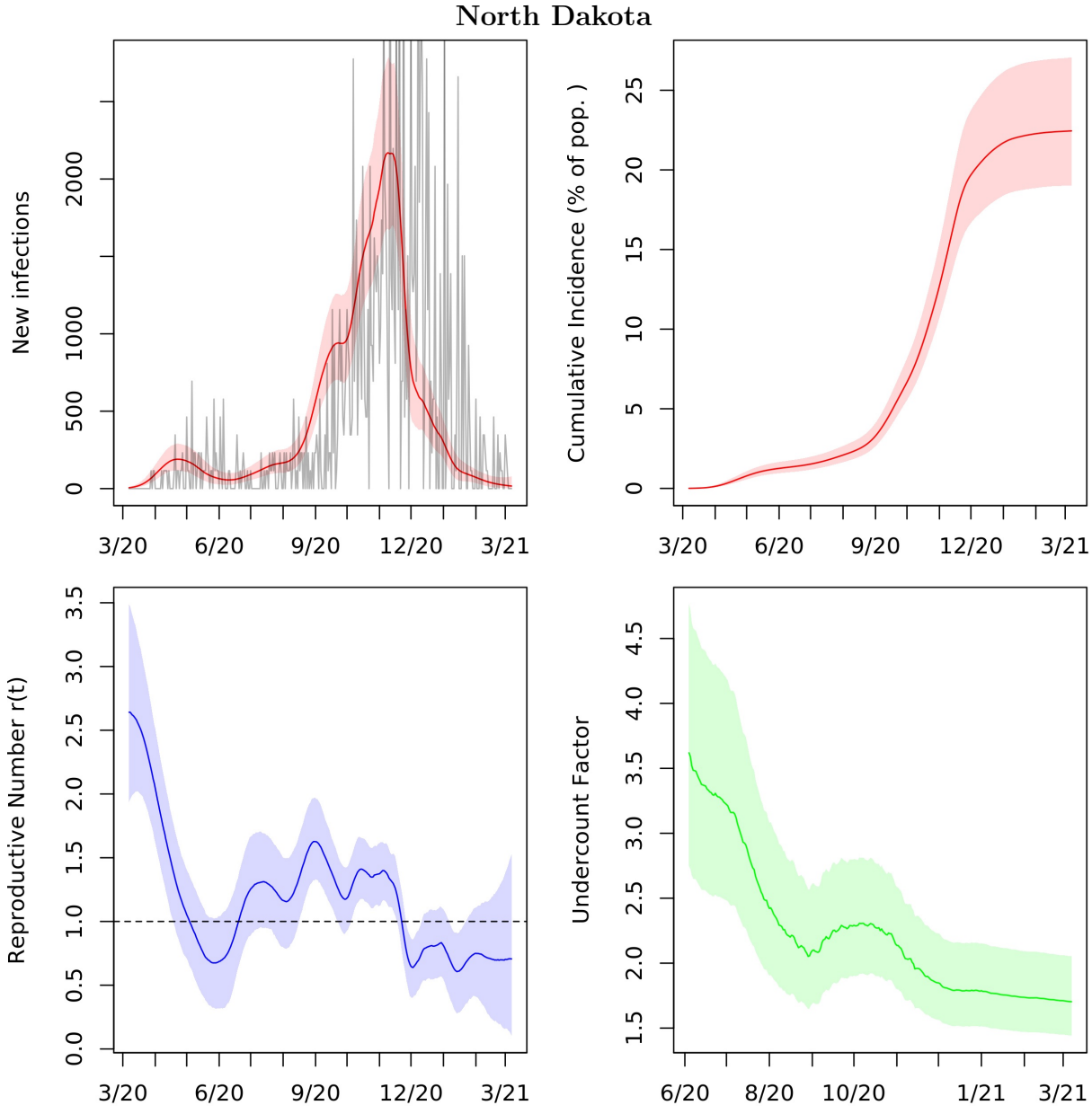


Figure A.29: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

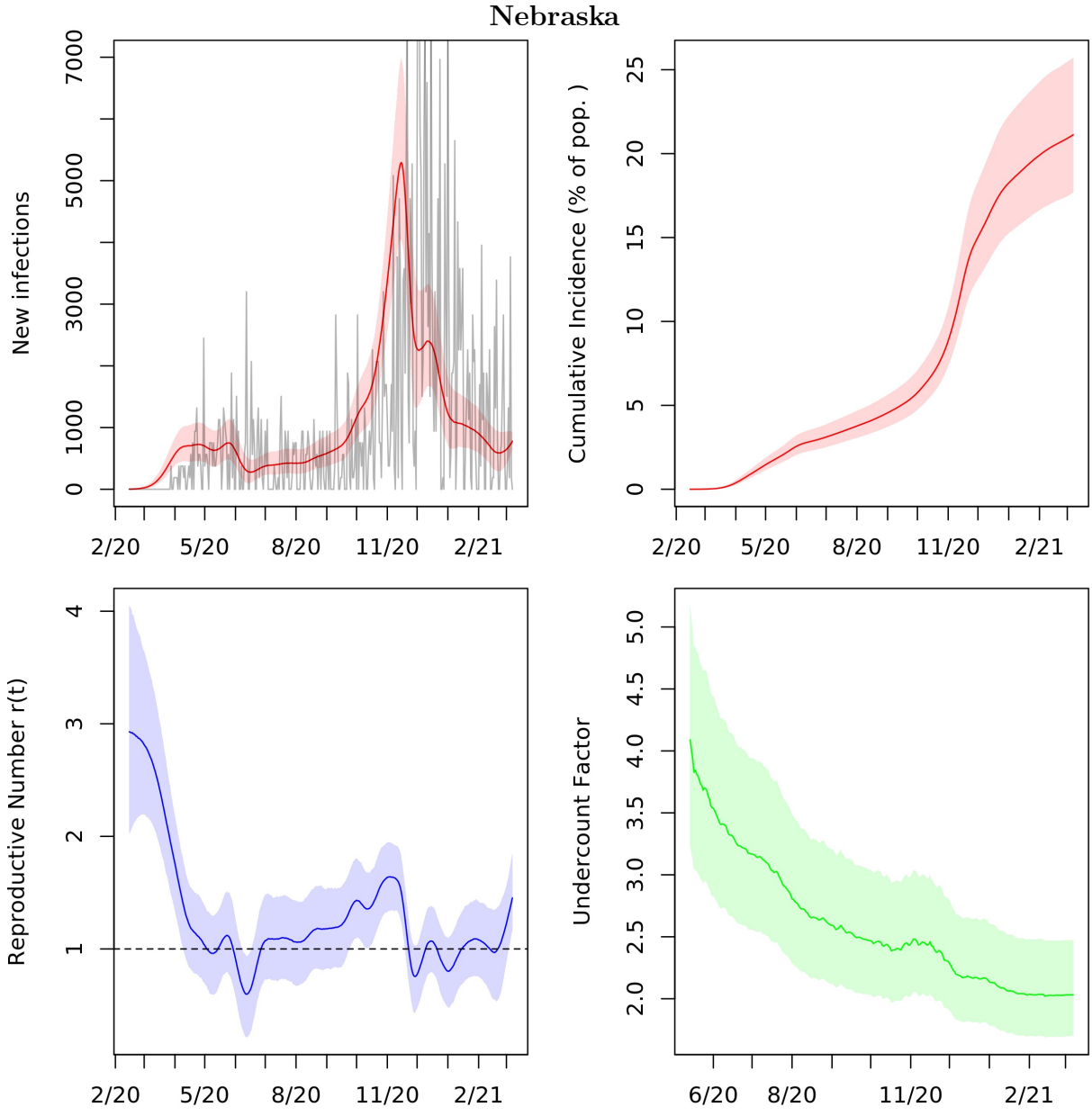


Figure A.30: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

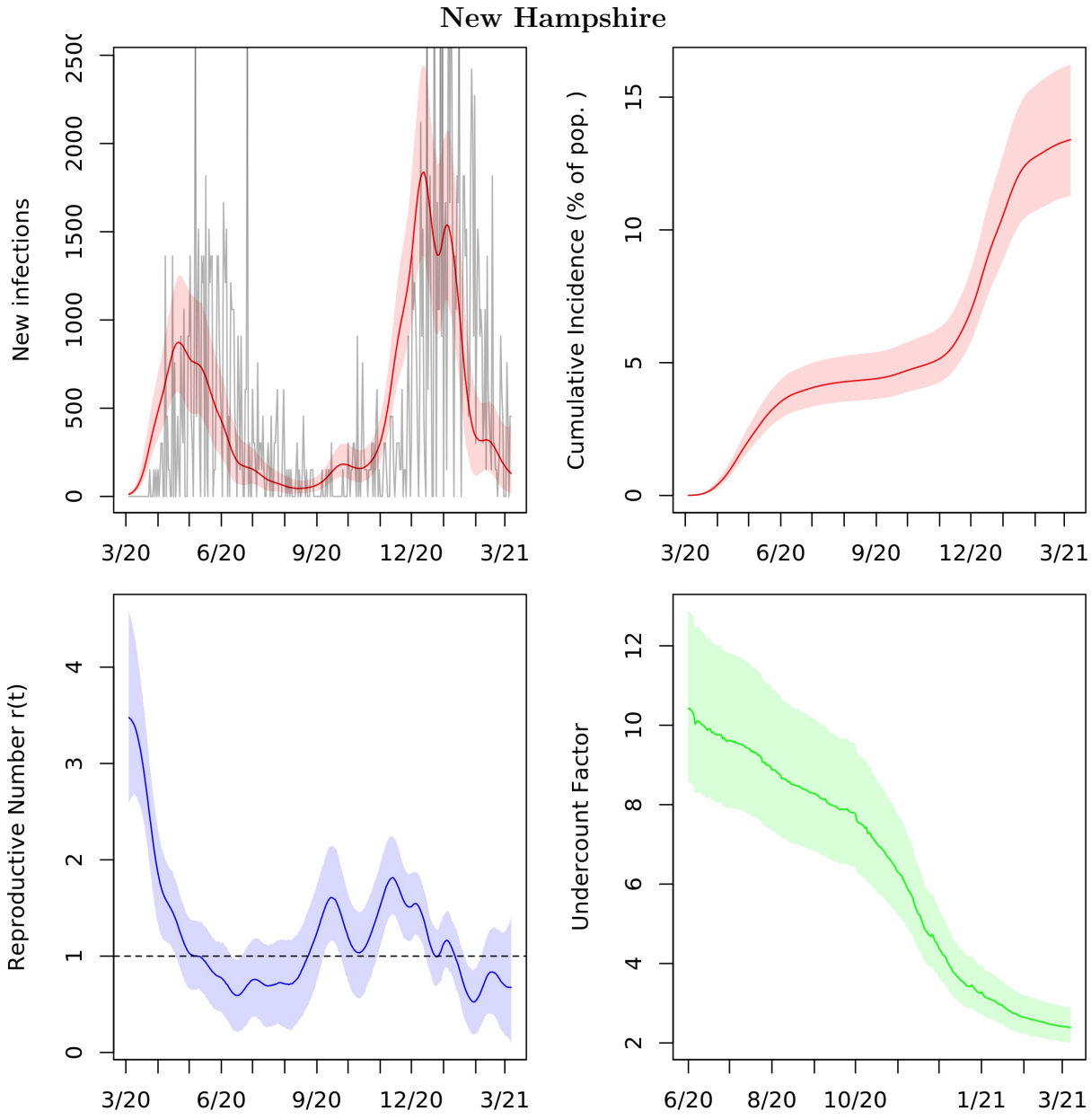


Figure A.31: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

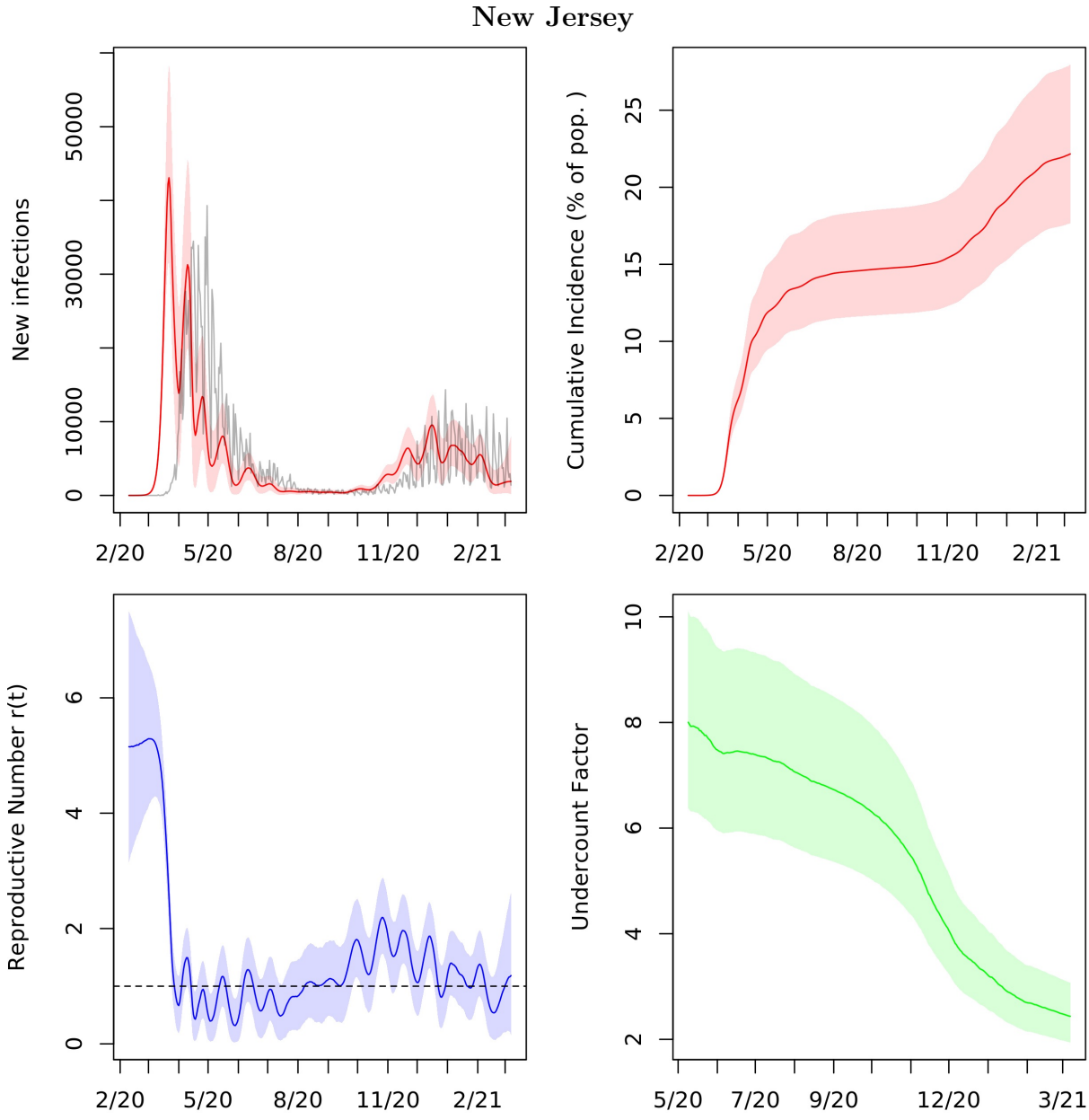


Figure A.32: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

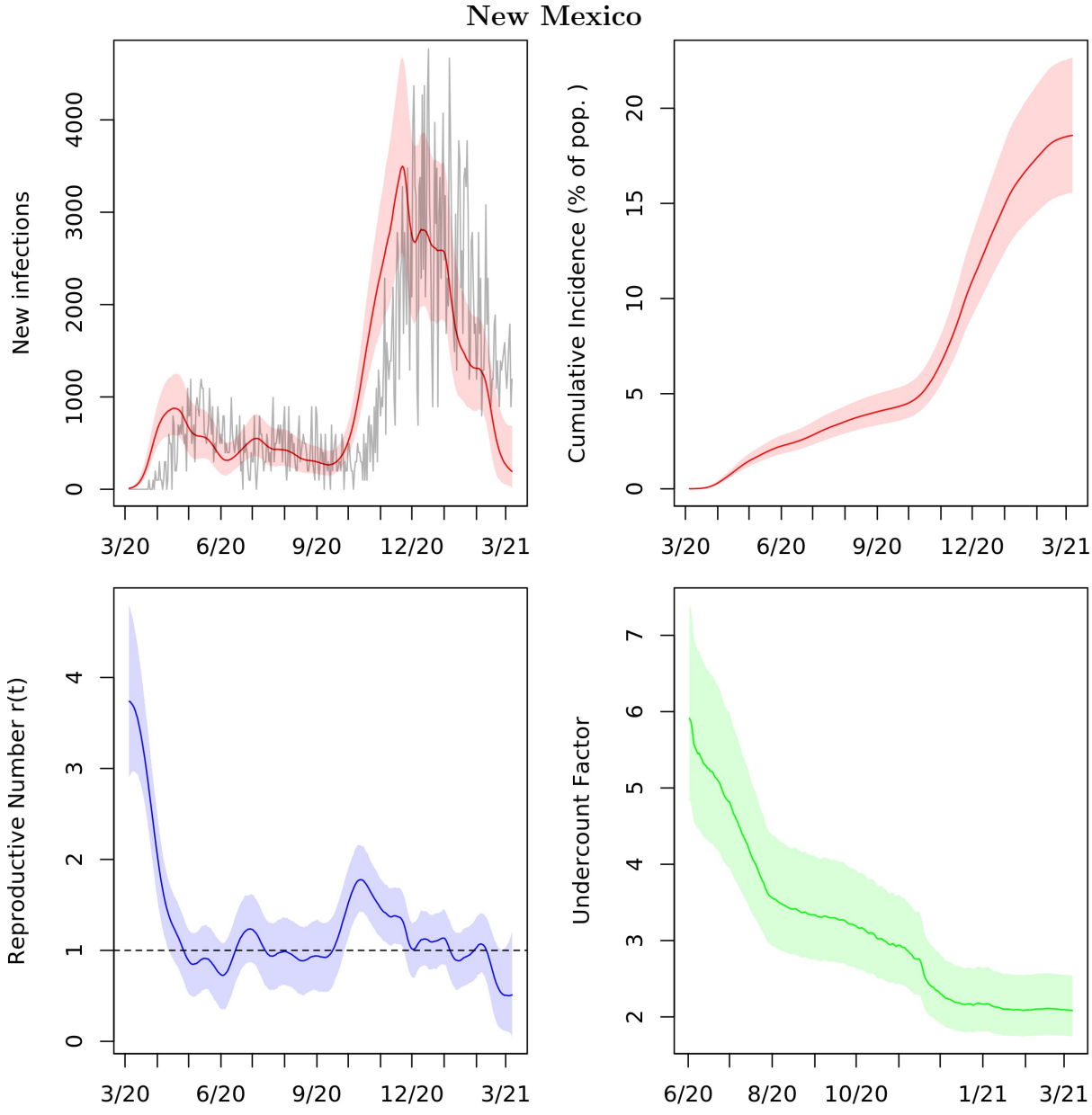


Figure A.33: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

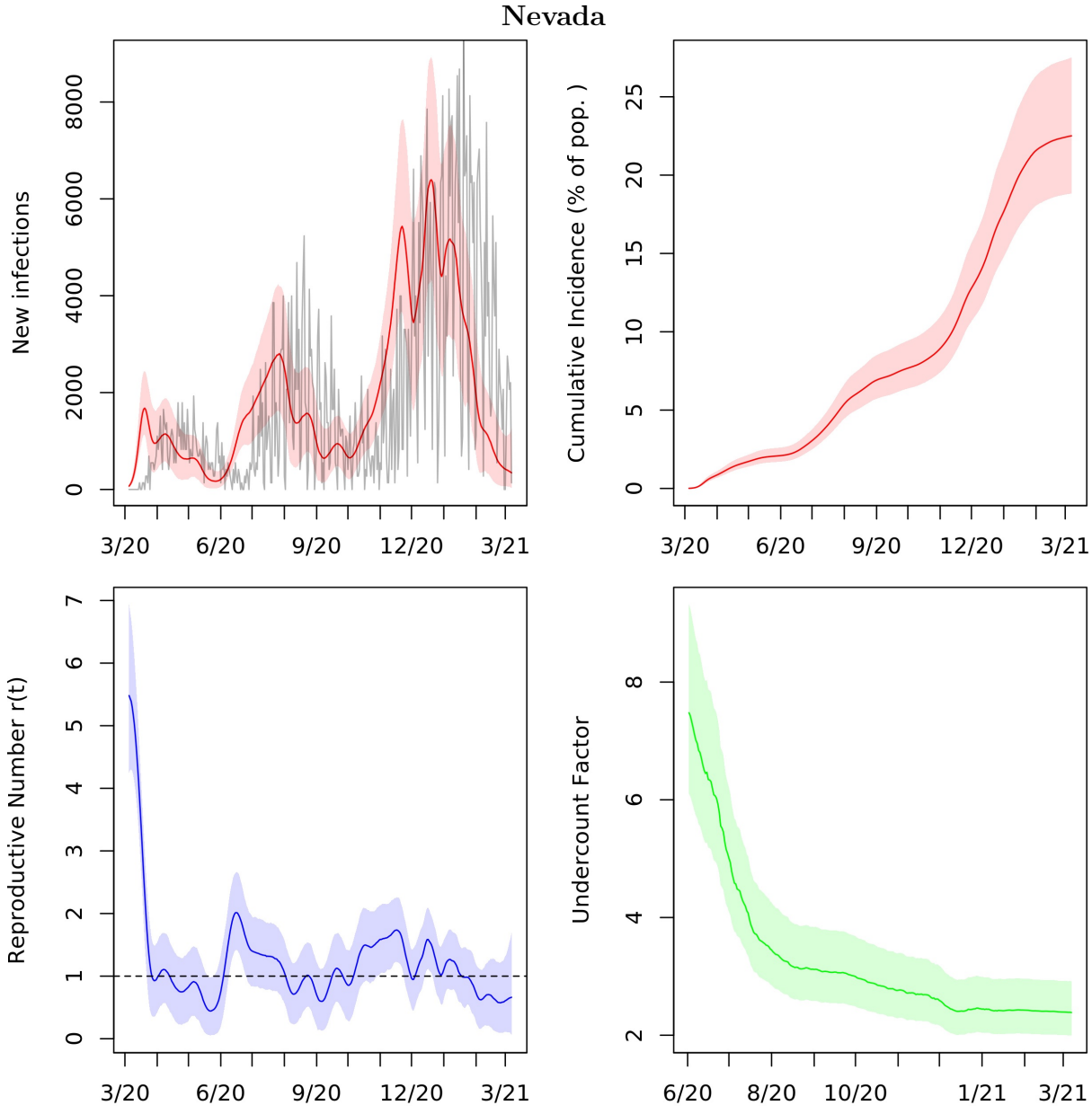


Figure A.34: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

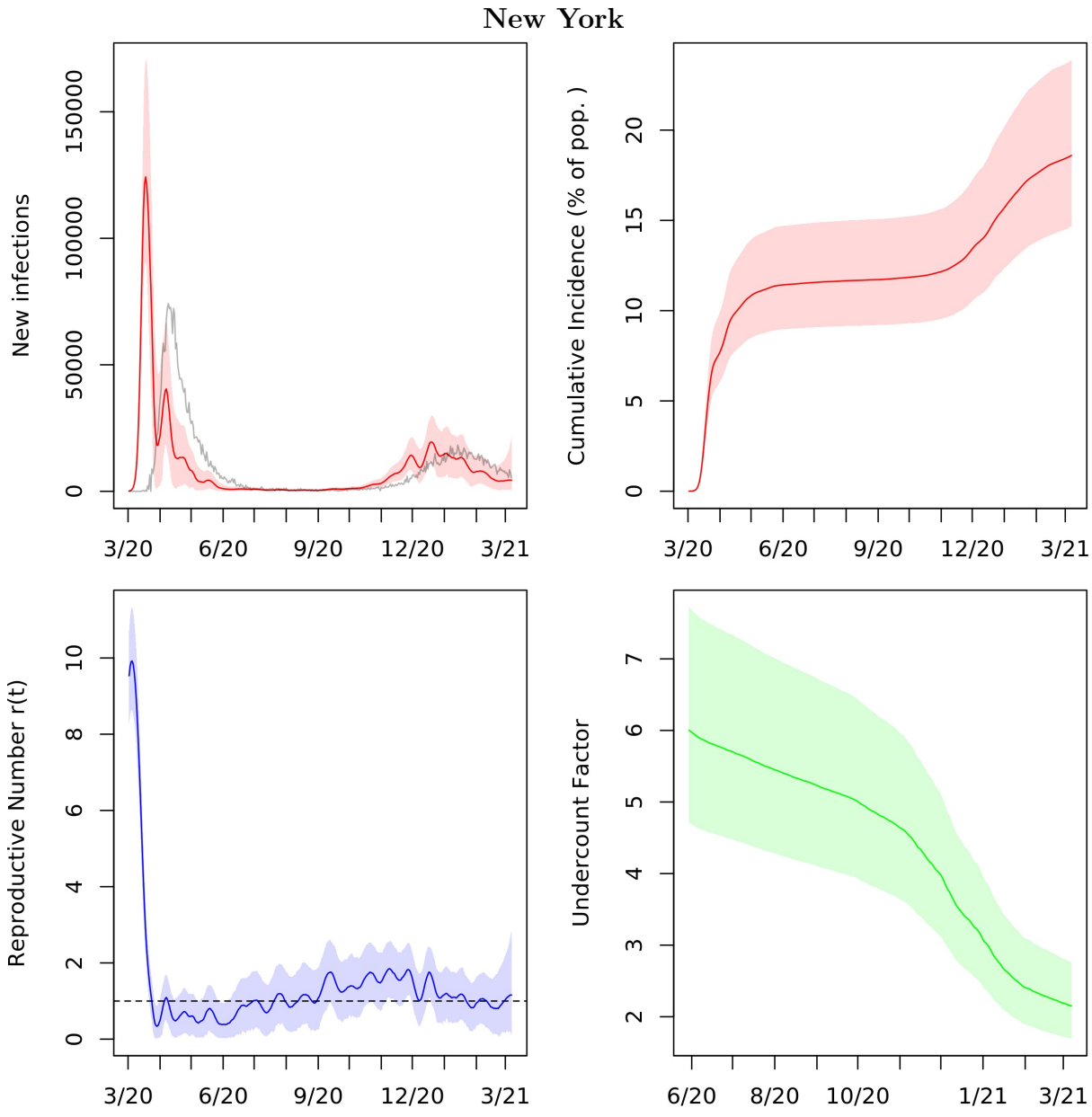


Figure A.35: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

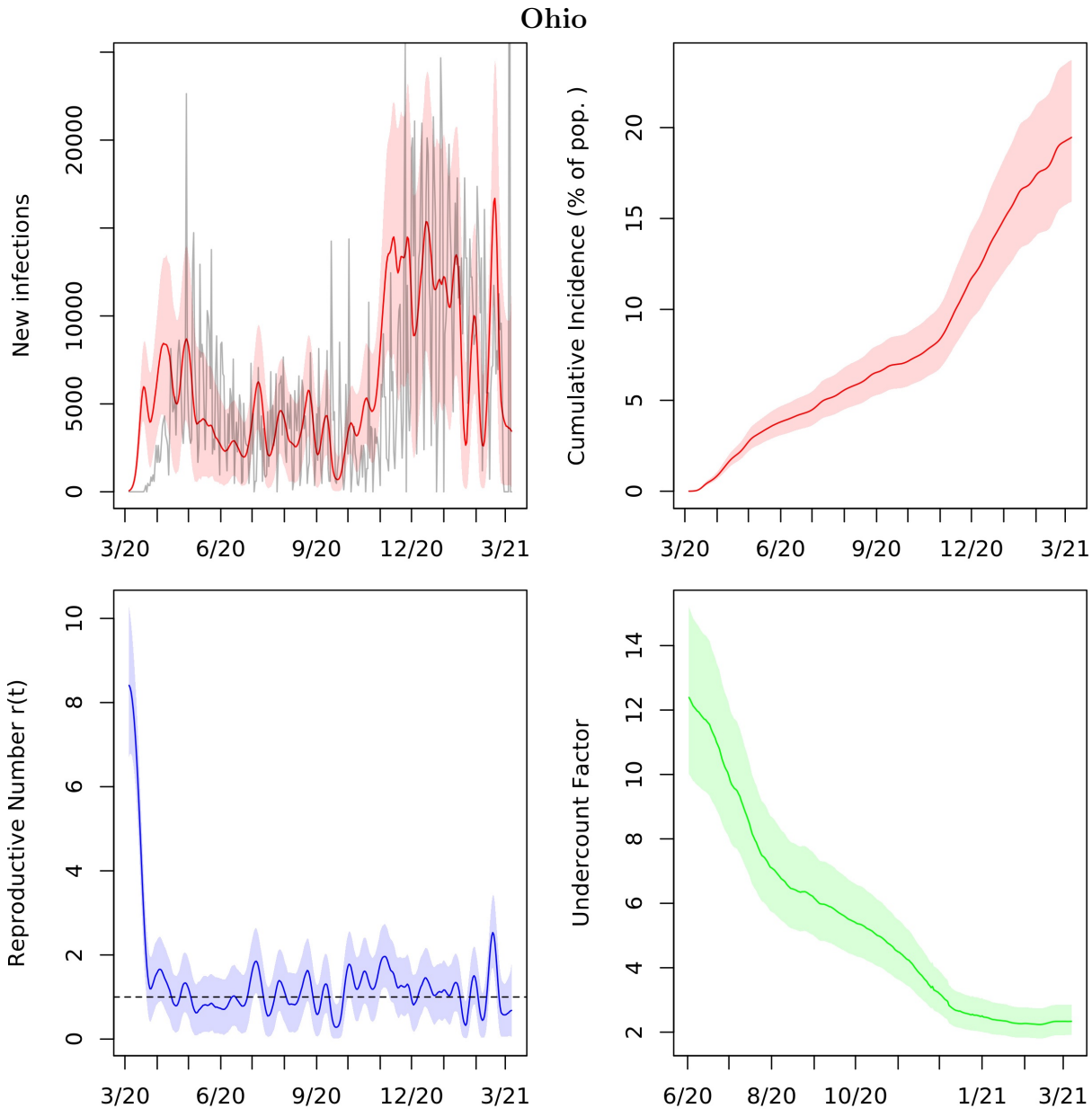


Figure A.36: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

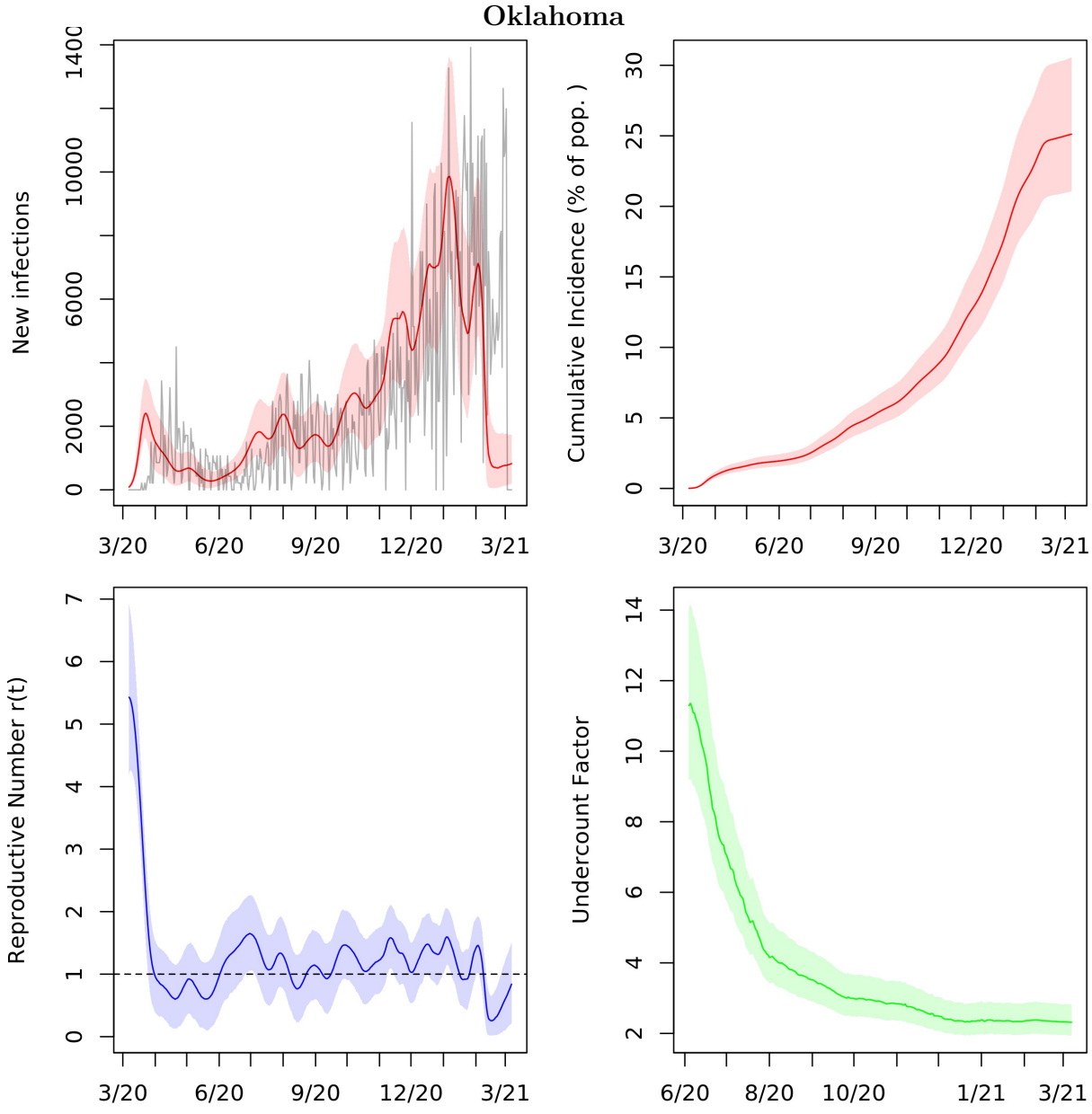


Figure A.37: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

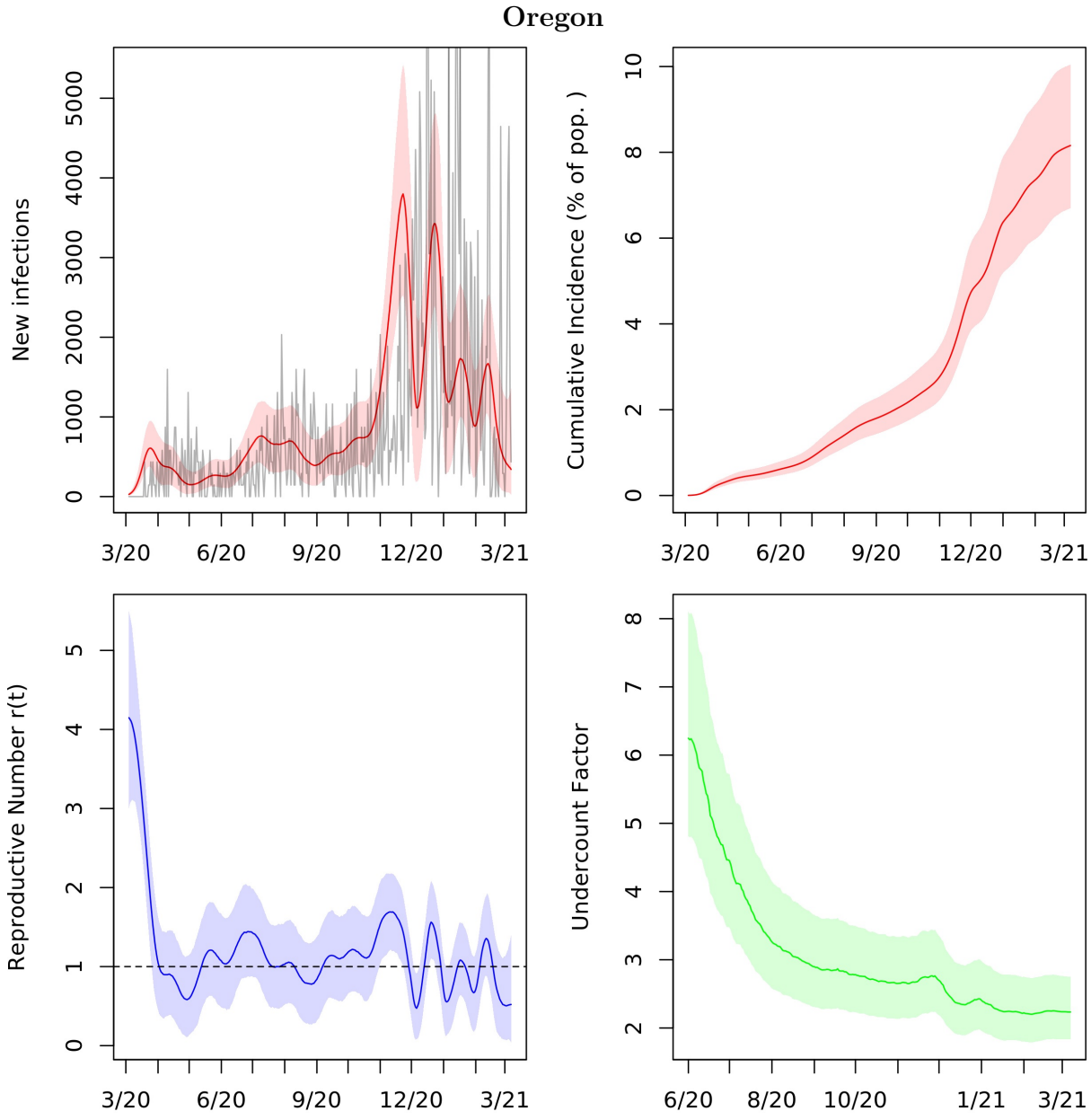


Figure A.38: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

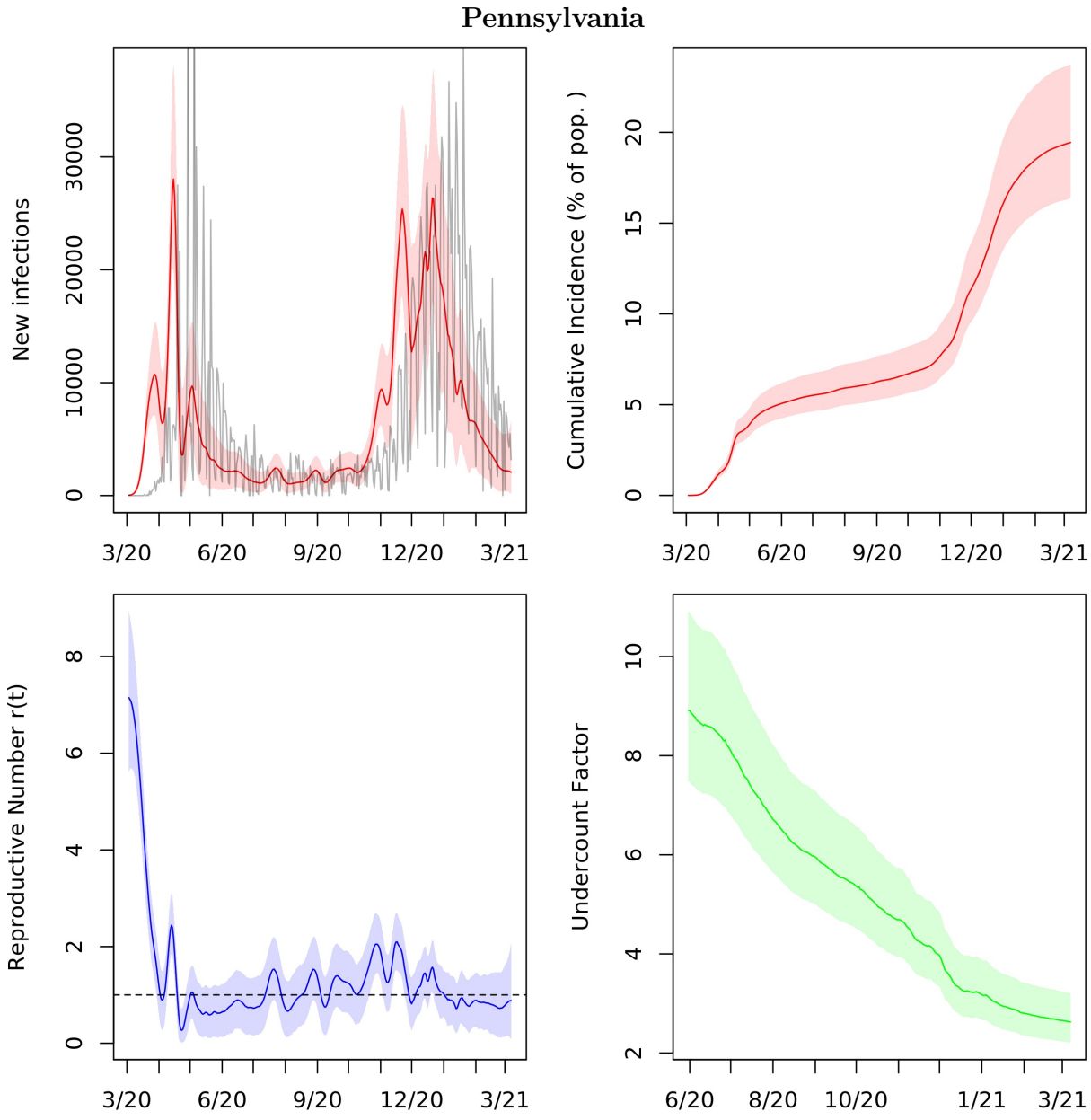


Figure A.39: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

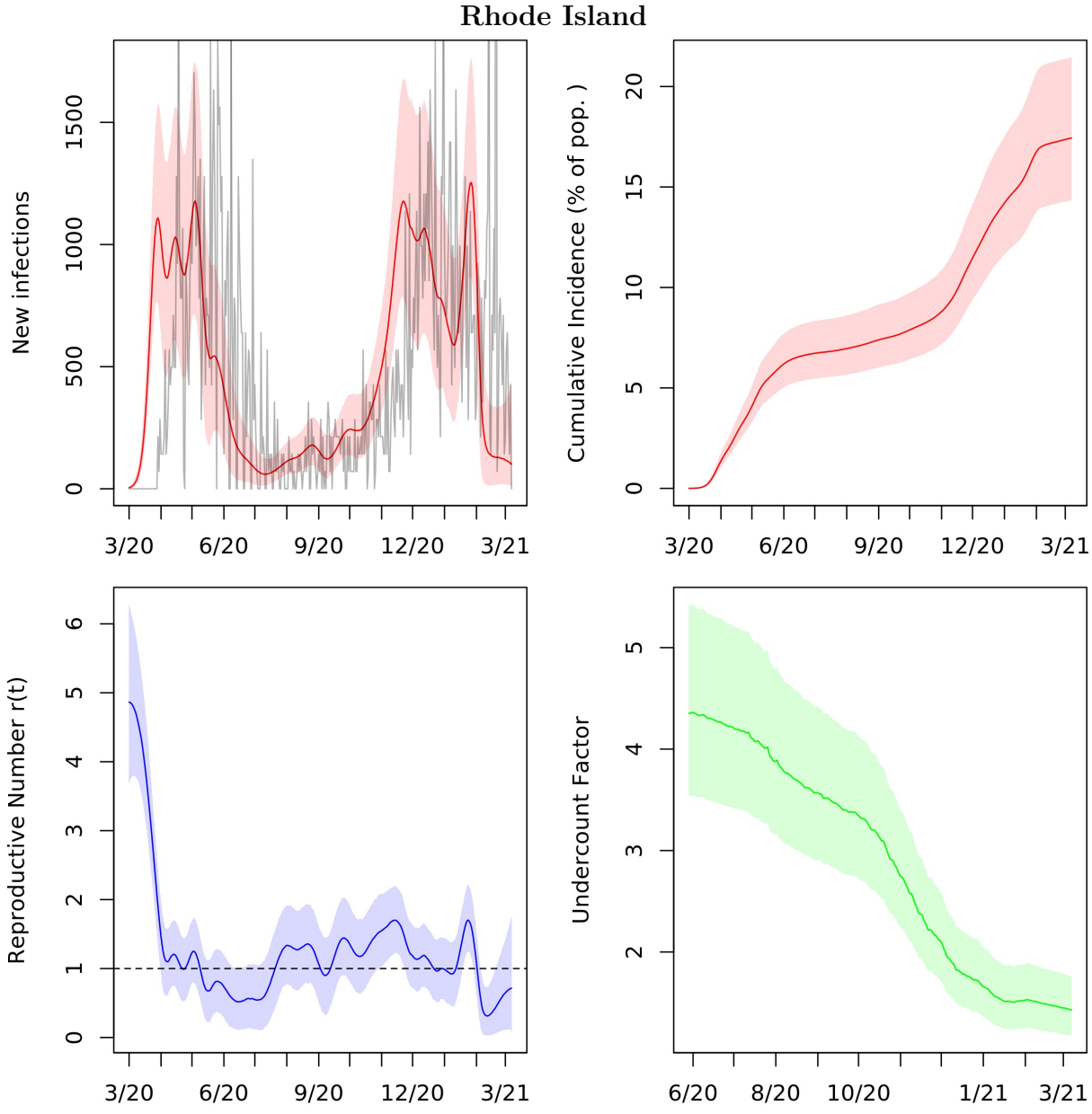


Figure A.40: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

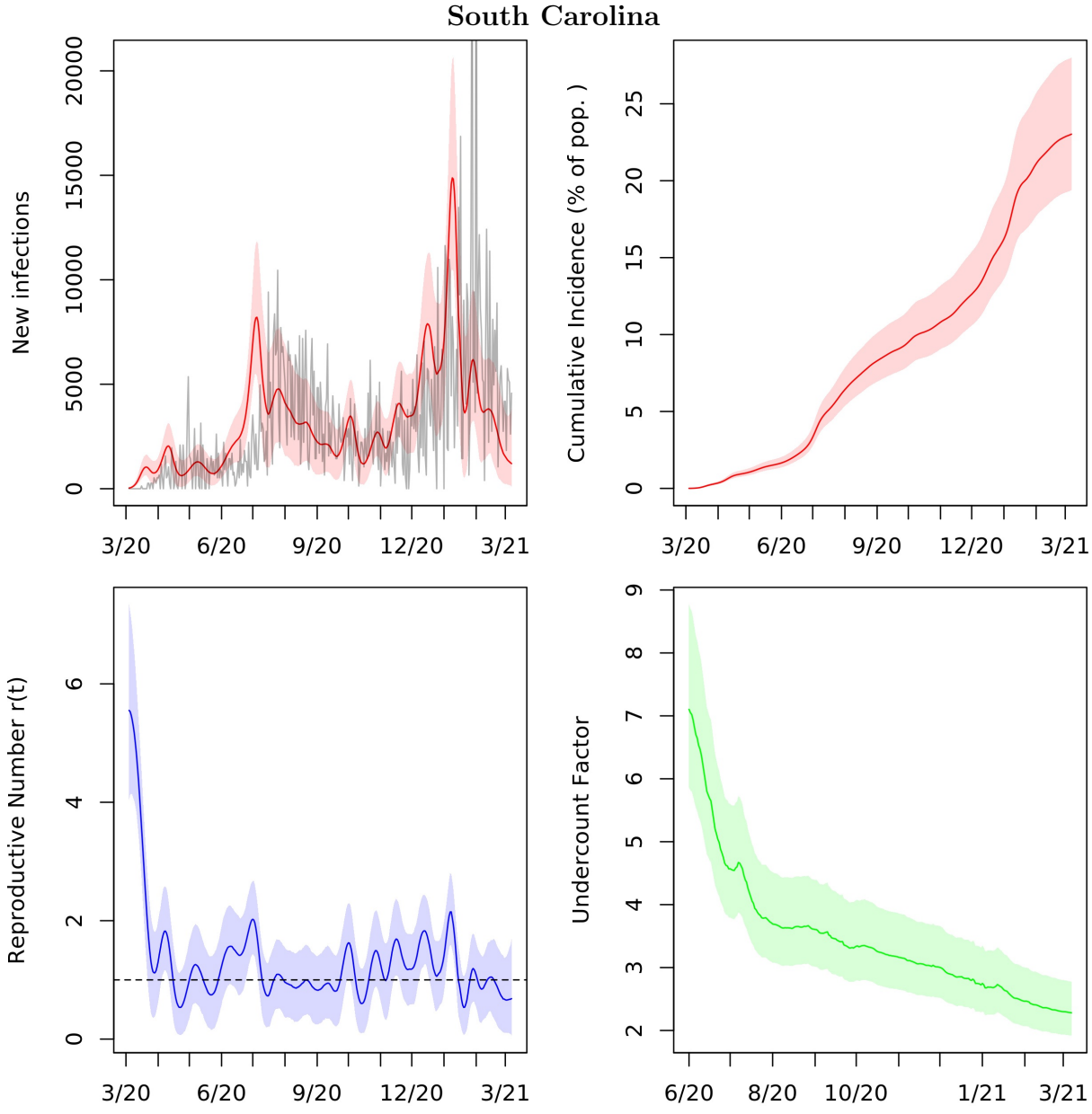


Figure A.41: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

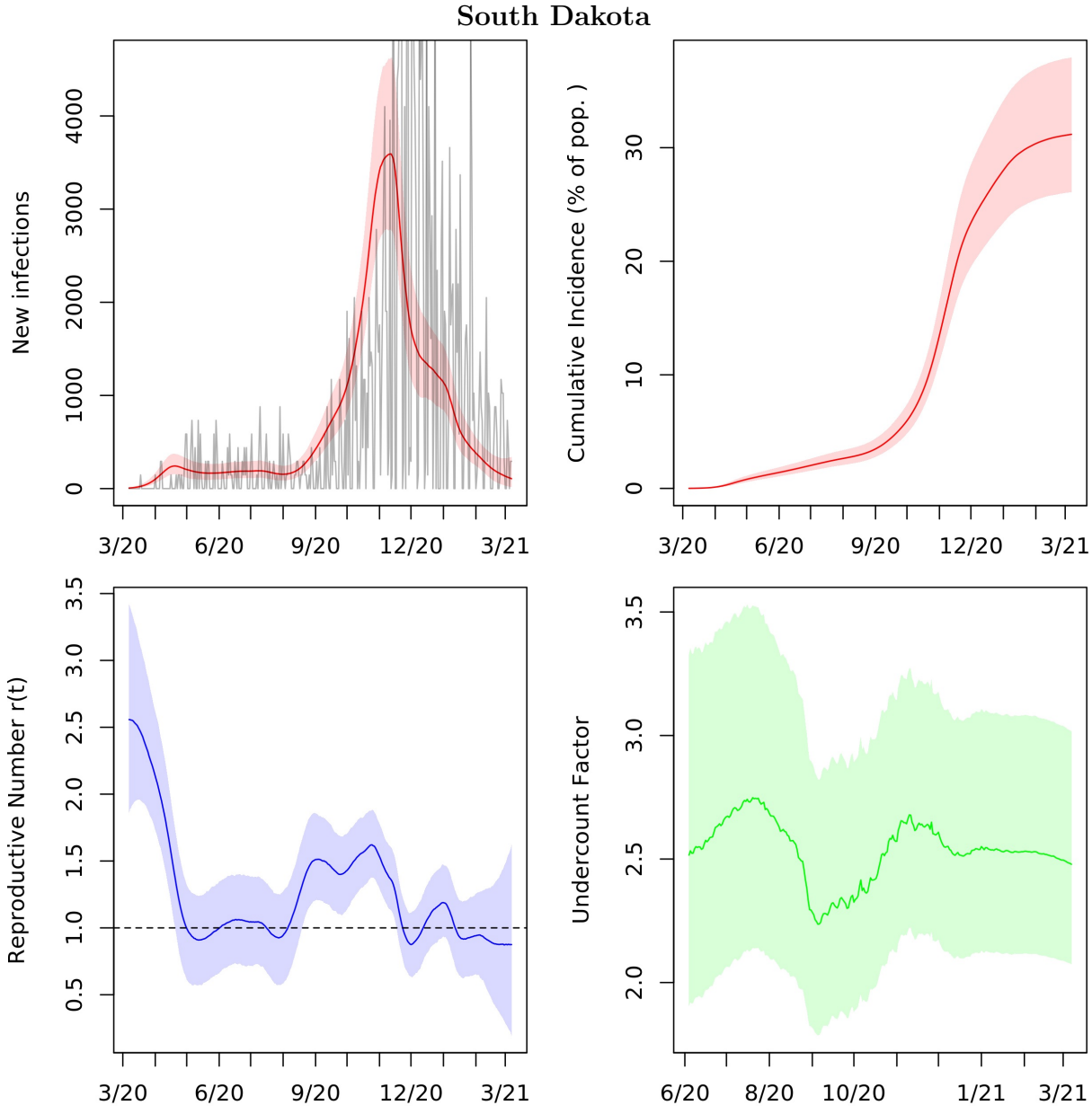


Figure A.42: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

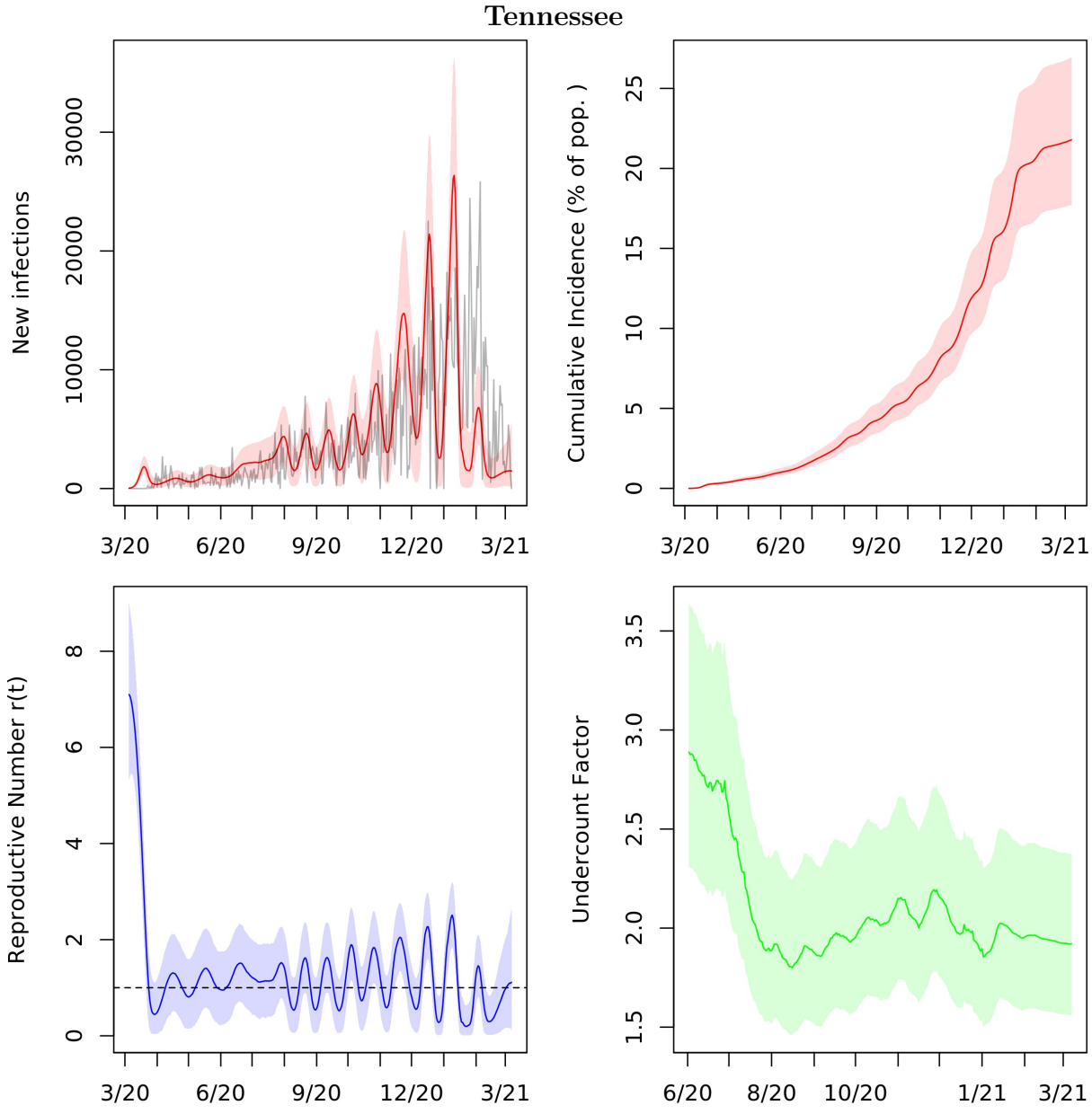


Figure A.43: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

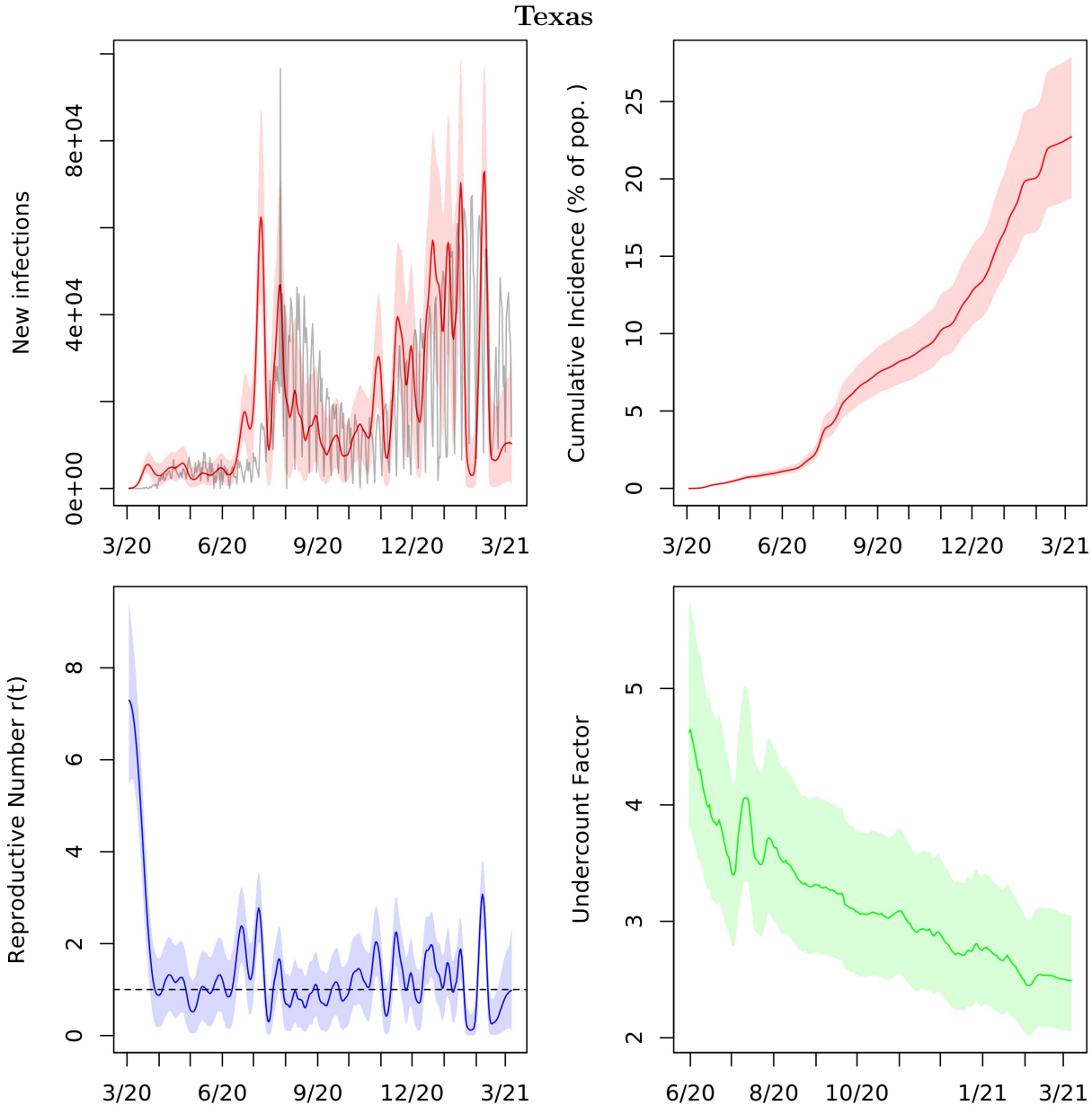


Figure A.44: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

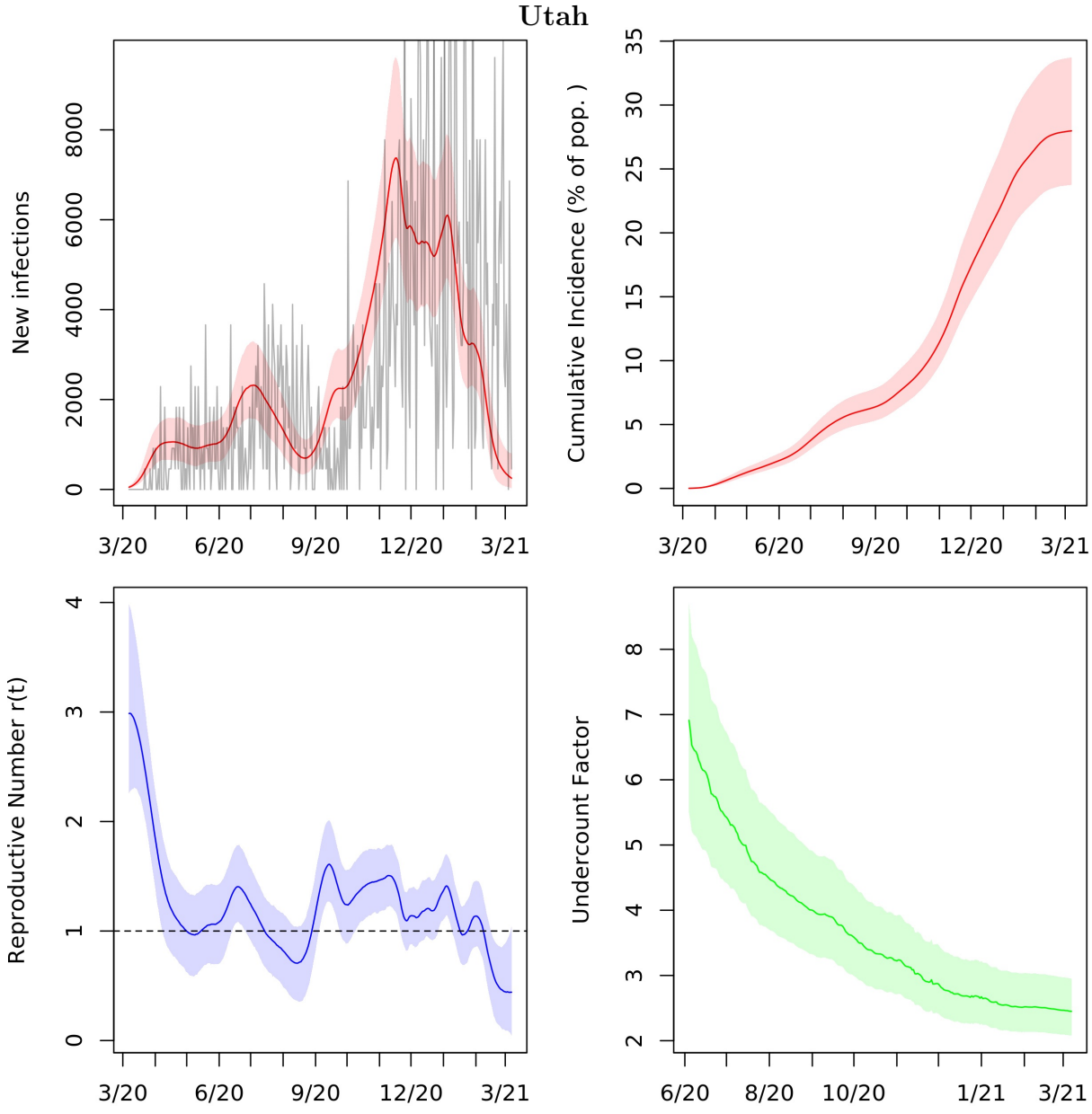


Figure A.45: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

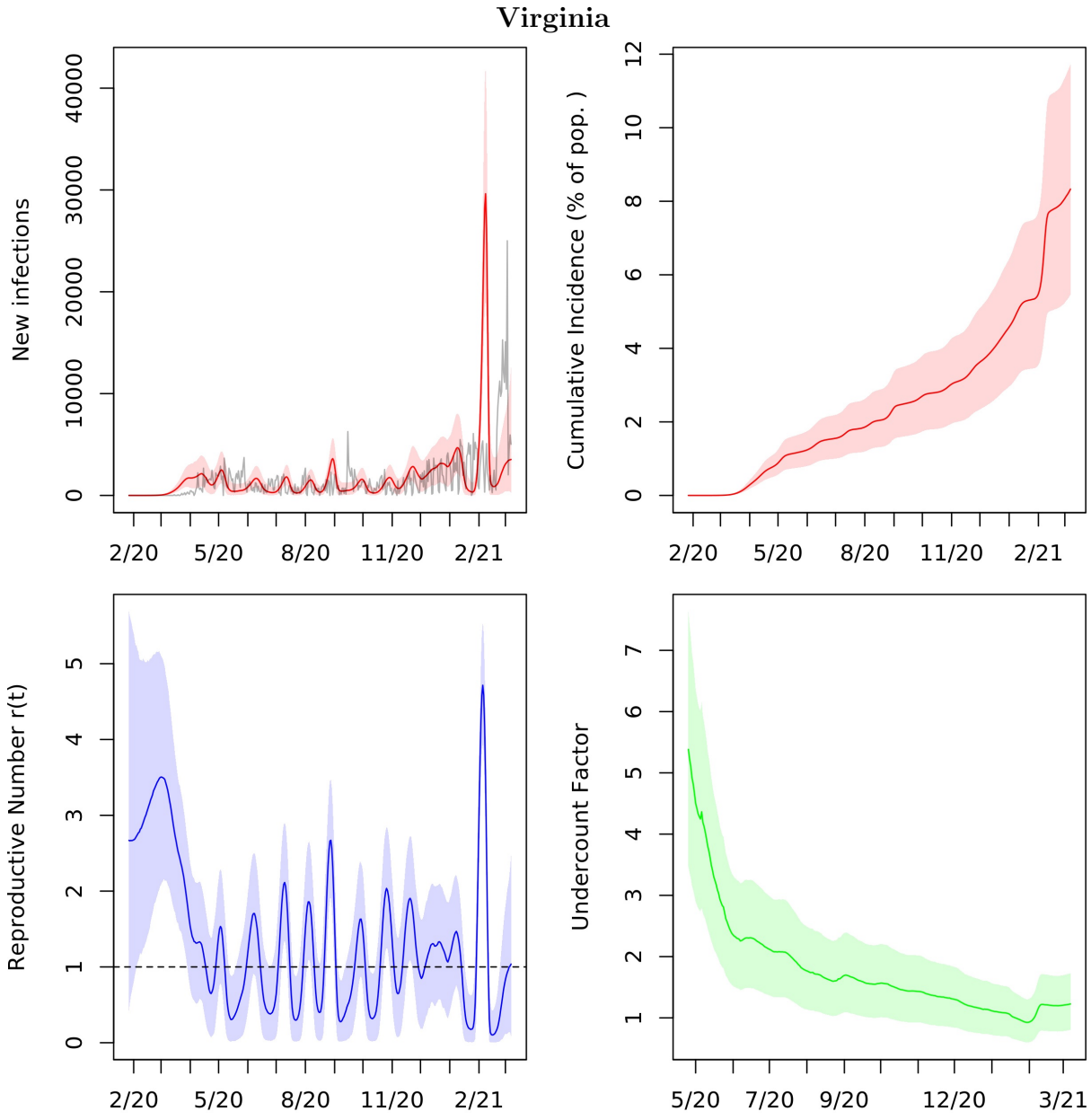


Figure A.46: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

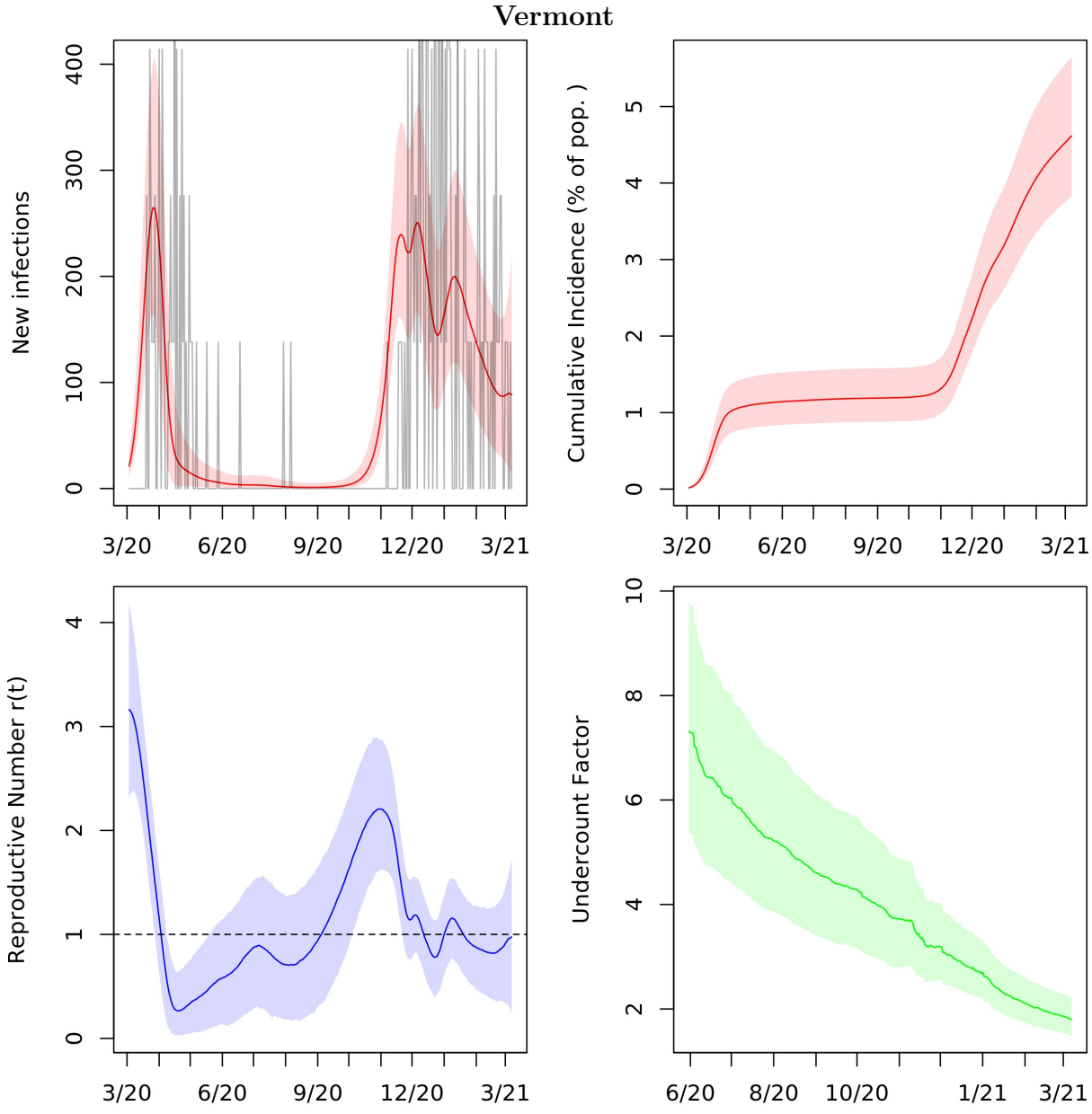


Figure A.47: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

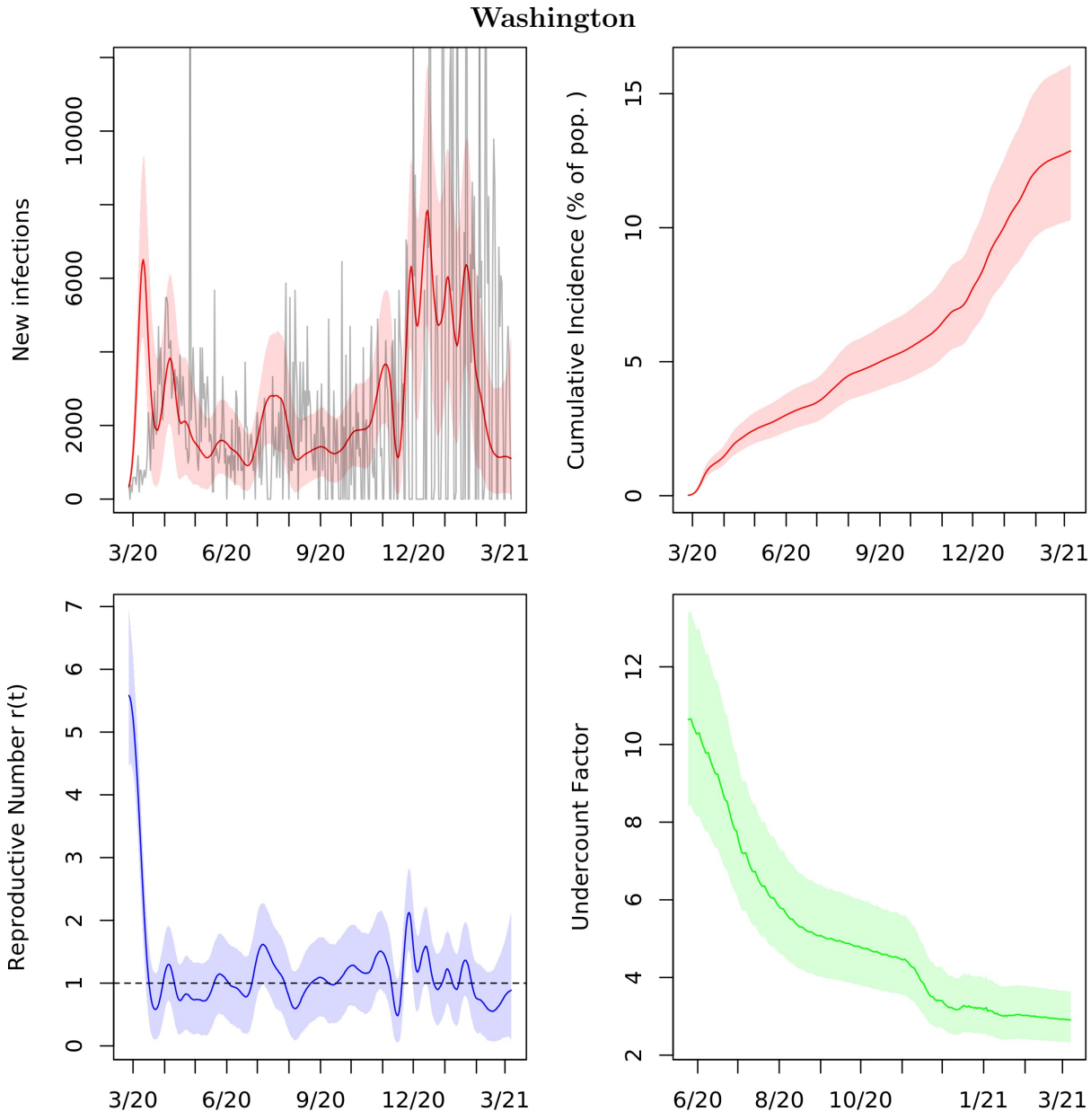


Figure A.48: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

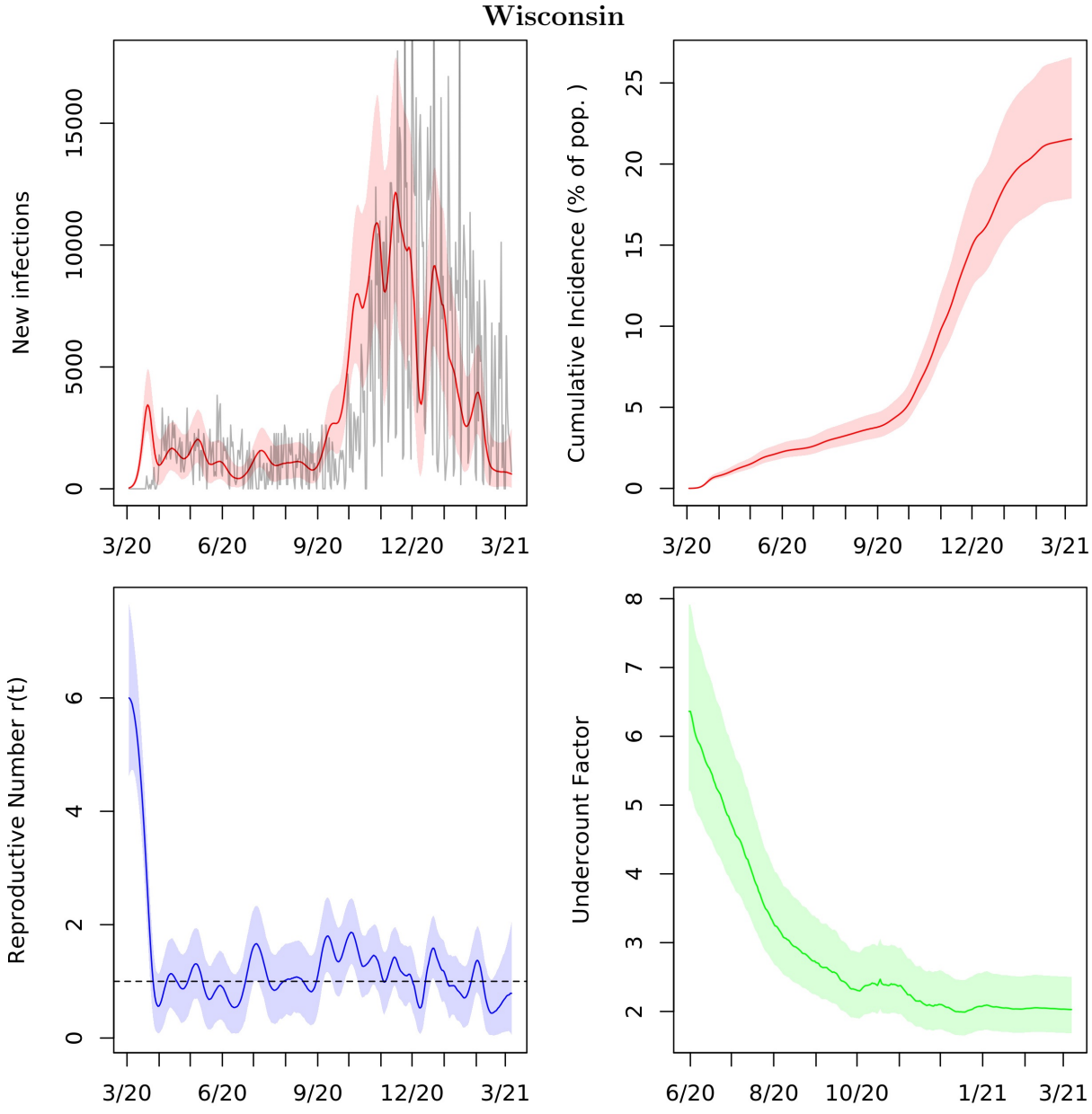


Figure A.49: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

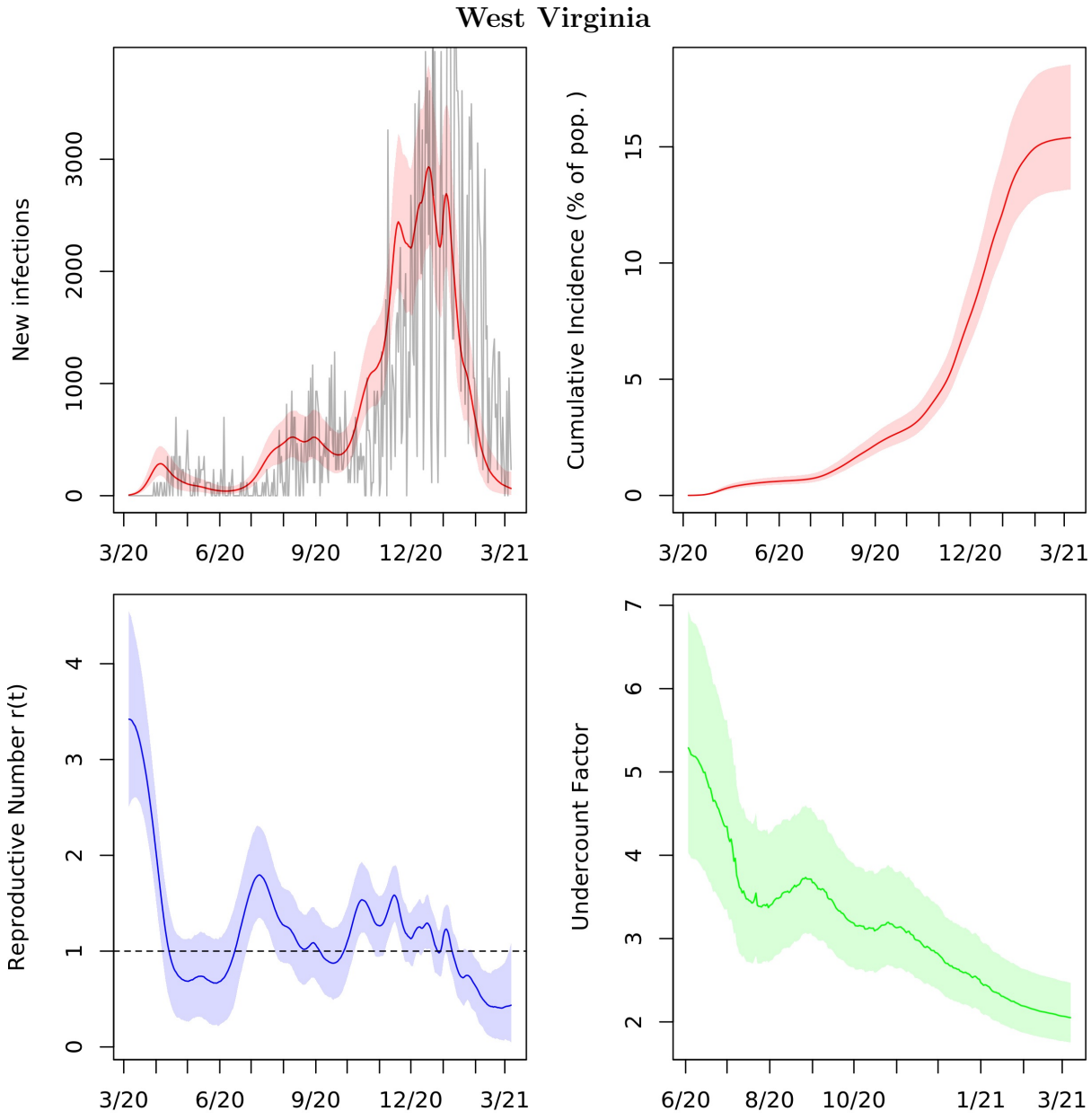


Figure A.50: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

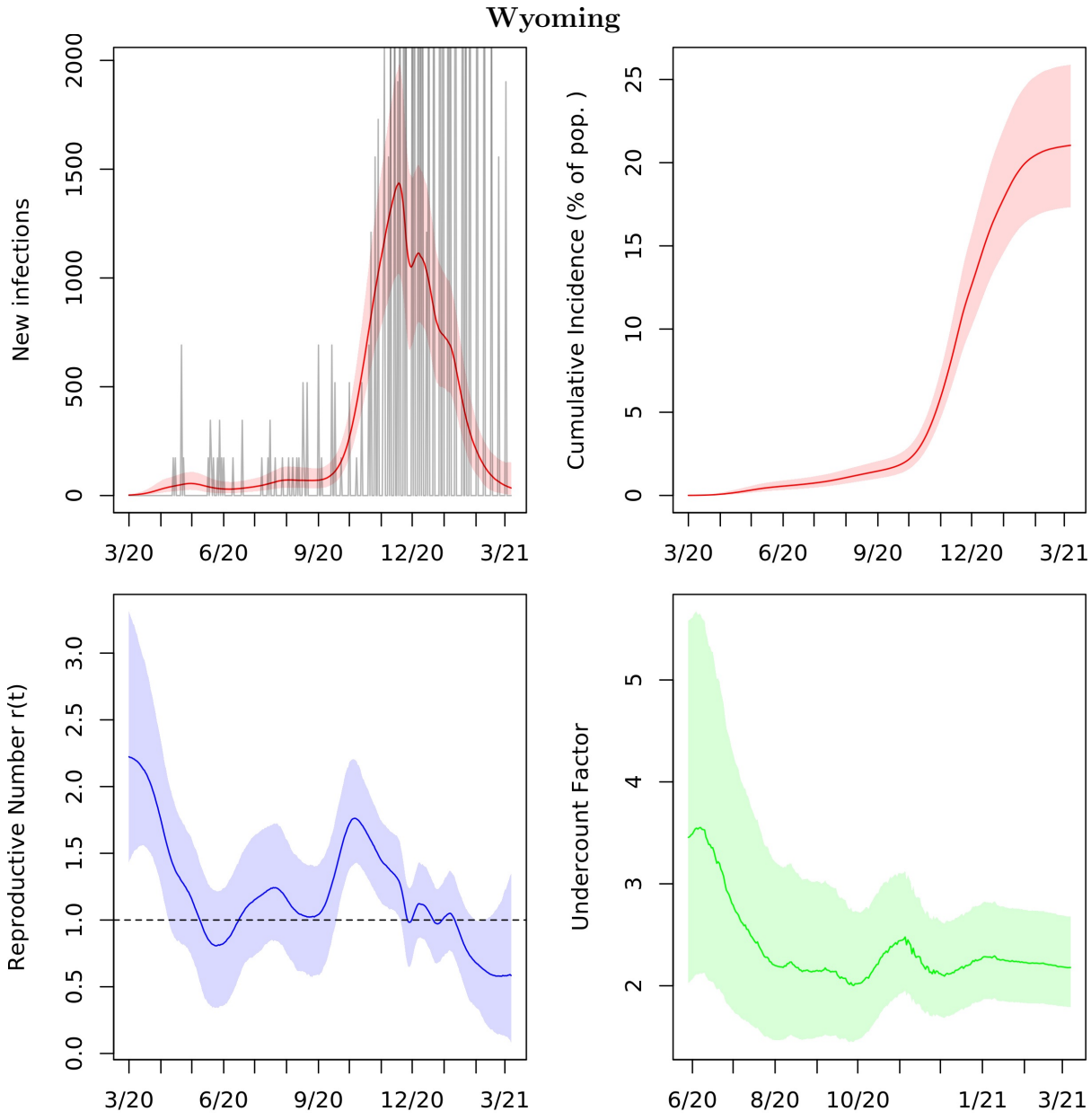


Figure A.51: Posterior median and middle 95% intervals for daily new infections, cumulative incidence, $r(t)$, and cumulative undercount from March 2020 to March 2021. In the top left panel, deaths divided by the posterior median IFR are plotted in grey for comparison.

A.2 BREASE posterior sampling

A.2.1 Sampling under monotonicity: no harm

Here we derive the BREASE posterior sampling algorithm under the "no harm" ($\eta_s = 0$) monotonicity model $M_{-}^{'}$ (A.12).

Theorem 4. *Let (θ_0, η_e) be random variables drawn according to Algorithm 3. Then (θ_0, η_e) are distributed according to the posterior of model $M_{-}^{'}$ (A.12).*

Proof. In this case, we make use of the posterior mixture representation

$$\pi(\theta_0, \eta_e|\mathcal{D}) = \sum_{P_1=0}^{N_1-y_1} \pi(\theta_0, \eta_e|P_1, \mathcal{D}) \times \pi(P_1|\mathcal{D}). \tag{A.1}$$

As discussed in Section 4.3.5, we have

$$P_1|(y_1, N_1, \theta_0, \eta_e) \sim \text{Binomial}\left(N_1 - y_1, \frac{\theta_0\eta_e}{1-\theta_1}\right).$$

Note that $\theta_1 = (1-\eta_e)\theta_0$ by hypothesis. Consequently, we have

$$\begin{aligned}
&\pi(\theta_0, \eta_e|P_1, \mathcal{D}) \\
&\quad \propto \pi(P_1, \mathcal{D}|\theta_0, \eta_e) \times \pi(\theta_0, \eta_e) \\
&\quad = \pi(P_1|\mathcal{D}, \theta_0, \eta_e) \times \pi(\mathcal{D}|\theta_0, \eta_e) \times \pi(\theta_0, \eta_e) \\
&\quad = \pi(P_1|y_1, N_1, \theta_0, \eta_e) \times \pi(\mathcal{D}|\theta_0, \eta_e) \times \pi(\theta_0, \eta_e) \\
&\quad = \text{Binomial}\left(P_1; N_1 - y_1, \frac{\theta_0\eta_e}{1-\theta_1}\right) \times \text{Binomial}(y_0; N_0, \theta_0) \times \text{Binomial}(y_1; N_1, \theta_1) \\
&\qquad \times \text{Beta}(\theta_0; \mu_0 n_0, (1-\mu_0)n_0) \times \text{Beta}(\eta_e; \mu_e n_e, (1-\mu_e)n_e) \\
&\quad \propto \theta_0^{y_0+y_1+P_1+\mu_0 n_0-1}(1-\theta_0)^{N-(y_0+y_1+P_1)+(1-\mu_0)n_0-1} \\
&\qquad \times \eta_e^{P_1+\mu_e n_e-1}(1-\eta_e)^{y_1+(1-\mu_e)n_e-1}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\pi(\theta_0, \eta_e | P_1, \mathcal{D}) \\
&\quad = \text{Beta}(\theta_0; y_0 + y_1 + P_1 + \mu_0 n_0, N - (y_0 + y_1 + P_1) + (1 - \mu_0) n_0) \\
&\quad\quad \times \text{Beta}(\eta_e; P_1 + \mu_e n_e, y_1 + (1 - \mu_e) n_e). \qquad \text{(A.2)}
\end{aligned}$$

Similarly, for the mixture weights we have

$$\begin{aligned}
\pi(P_1|\mathcal{D}) &= \int \pi(P_1, \theta_0, \eta_e | \mathcal{D}) d\theta_0 d\eta_e \\
&= \int \pi(P_1 | \theta_0, \eta_e, \mathcal{D}) \pi(\theta_0, \eta_e | \mathcal{D}) d\theta_0 d\eta_e \\
&\propto \binom{N_1 - y_1}{P_1} \text{B}(P_1 + \mu_e n_e, y_1 + (1 - \mu_e) n_e) \\
&\quad \times \text{B}(y_0 + y_1 + P_1 + \mu_0 n_0, N - (y_0 + y_1 + P_1) + (1 - \mu_0) n_0). \qquad \text{(A.3)}
\end{aligned}$$

Algorithm 3 defines the procedure to sample from the distribution A.1 based on these calculations. Algorithm 4 defines the corresponding data-augmented Gibbs sampler. □

Algorithm 3 "No harm" ($\eta_s = 0$) exact posterior sampling algorithm

---

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters (μ_0, μ_e, n_0, n_e), and desired number of posterior samples T.

Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $P_1 \in \{0, \ldots, N_1 - y_1\}$ with probability $\pi(P_1|\mathcal{D})$ given by (A.3).

(ii) Sample (θ_0, η_e) conditional on $(P_1, \mathcal{D})$ from the independent beta distribution (A.2).

Output: Posterior samples $\{(\theta_0^{(t)}, \eta_e^{(t)})\}_{t \in \{1,\ldots,T\}}$.

Algorithm 4 "No harm" ($\eta_s = 0$) data-augmentation algorithm

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters (μ_0, μ_e, n_0, n_e), desired number of posterior samples T, number of burn-in iterations B, and parameter initialization $(\theta_0^{(0)}, \eta_e^{(0)}) \in (0,1)^2$.

Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $P_1^{(t)}$ conditional on $(\theta_0^{(t-1)}, \eta_e^{(t-1)}, \mathcal{D})$ from the binomial distribution

$$P_1^{(t)} \sim \text{Binomial}\left(N_1 - y_1, \frac{\theta_0^{(t-1)}\eta_e^{(t-1)}}{1 - \theta_1^{(t-1)}}\right), \tag{A.4}$$

where $\theta_1^{(t-1)} = \theta_0^{(t-1)}(1 - \eta_e^{(t-1)})$.

(ii) Sample $(\theta_0^{(t)}, \eta_e^{(t)})$ conditional on $(P_1^{(t)}, \mathcal{D})$ from the independent beta (A.2).

Output: Posterior samples after burn-in $\{(\theta_0^{(t)}, \eta_e^{(t)})\}_{t \in \{B+1, \ldots, T\}}$.

A.2.2 Sampling under monotonicity: no benefit

Here we derive the BREASE posterior sampling algorithm under the "no benefit" ($\eta_e = 0$) monotonicity model $M_+^{'}$ (A.13).

Theorem 5. *Let (θ_0, η_s) be random variables drawn according to Algorithm 5. Then (θ_0, η_s) are distributed according to the posterior of model $M_+^{'}$ (A.13).*

Proof. In this case, we make use of the posterior mixture representation

$$\pi(\theta_0, \eta_s | \mathcal{D}) = \sum_{C_1=0}^{y_1} \pi(\theta_0, \eta_s | C_1, \mathcal{D}) \times \pi(C_1 | \mathcal{D}). \tag{A.5}$$

As discussed in Section 4.3.5, we have

$$C_1 | (y_1, \theta_0, \eta_s) \sim \text{Binomial}\left(y_1, \frac{(1 - \theta_0)\eta_s}{\theta_1}\right).$$

Note that $\theta_1 = \theta_0 + (1-\theta_0)\eta_s$ by hypothesis. Consequently, we have

$$\begin{aligned}
&\pi(\theta_0, \eta_s | C_1, \mathcal{D}) \\
&\quad \propto \pi(C_1, \mathcal{D} | \theta_0, \eta_s) \times \pi(\theta_0, \eta_s) \\
&\quad = \pi(C_1 | \mathcal{D}, \theta_0, \eta_s) \times \pi(\mathcal{D} | \theta_0, \eta_s) \times \pi(\theta_0, \eta_s) \\
&\quad = \pi(C_1 | y_1, \theta_0, \eta_s) \times \pi(\mathcal{D} | \theta_0, \eta_s) \times \pi(\theta_0, \eta_s) \\
&\quad = \text{Binomial}\left(C_1; y_1, \frac{(1-\theta_0)\eta_s}{\theta_1}\right) \times \text{Binomial}(y_0; N_0, \theta_0) \times \text{Binomial}(y_1; N_1, \theta_1) \\
&\qquad \times \text{Beta}(\theta_0; \mu_0 n_0, (1-\mu_0) n_0) \times \text{Beta}(\eta_s; \mu_s n_s, (1-\mu_s) n_s) \\
&\quad \propto \theta_0^{y_0 + y_1 - C_1 + \mu_0 n_0 - 1} (1-\theta_0)^{N - (y_0 + y_1 - C_1) + (1-\mu_0) n_0 - 1} \\
&\qquad \times \eta_s^{C_1 + \mu_s n_s - 1} (1-\eta_s)^{N_1 - y_1 + (1-\mu_s) n_s - 1}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\pi(\theta_0, \eta_s | C_1, \mathcal{D}) \\
&\qquad = \text{Beta}(\theta_0; y_0 + y_1 - C_1 + \mu_0 n_0, N - (y_0 + y_1 - C_1) + (1-\mu_0) n_0) \\
&\qquad\quad \times \text{Beta}(\eta_s; C_1 + \mu_s n_s, N_1 - y_1 + (1-\mu_s) n_s). \qquad \text{(A.6)}
\end{aligned}$$

Similarly, for the mixture weights we have

$$\begin{aligned}
\pi(C_1 | \mathcal{D}) &= \int \pi(C_1, \theta_0, \eta_s | \mathcal{D}) d\theta_0 d\eta_s \\
&= \int \pi(C_1 | \theta_0, \eta_s, \mathcal{D}) \pi(\theta_0, \eta_s | \mathcal{D}) d\theta_0 d\eta_s \\
&\propto \binom{y_1}{C_1} \text{B}(y_0 + y_1 - C_1 + \mu_0 n_0, N - (y_0 + y_1 - C_1) + (1-\mu_0) n_0) \\
&\qquad \times \text{B}(C_1 + \mu_s n_s, N_1 - y_1 + (1-\mu_s) n_s). \qquad \text{(A.7)}
\end{aligned}$$

Algorithm 5 defines the procedure to sample from the distribution A.5 based on these calculations. Algorithm 6 defines the corresponding data-augmented Gibbs sampler. □

Algorithm 5 "No benefit" ($\eta_e = 0$) exact posterior sampling algorithm

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters (μ_0, μ_s, n_0, n_s), and desired number of posterior samples T.

Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $C_1 \in \{0, \ldots, y_1\}$ conditional on $\mathcal{D}$ with probability $\pi(C_1|\mathcal{D})$ given by (A.7).

(ii) Sample (θ_0, η_s) conditional on $(C_1, \mathcal{D})$ from the independent beta distribution (A.6).

Output: Posterior samples $\{(\theta_0^{(t)}, \eta_s^{(t)})\}_{t\in\{1,\ldots,T\}}$.

A.2.3 Sampling with an alternate prior under $H_0 : \theta_0 = \theta_1$

We now derive a sampling algorithm for the aggregated Dirichlet prior under H_0 introduced in Section A.3.1:

$$\mathbf{p}^* = (p_{00}, p_{10}^*, p_{11}) \sim \text{Dirichlet}((1-\mu_s)n_s, \mu_e n_e + \mu_s n_s, (1-\mu_e)n_e), \qquad p_{10}^* = p_{10} + p_{01}.$$

The algorithm is based on the posterior decomposition

$$\pi(\mathbf{p}^*|\mathcal{D}) = \sum_{w(0)=0}^{y_0+y_1} \sum_{w(1)=0}^{N_0+N_1-y_0-y_1} \pi(\mathbf{p}^*|w(0), w(1), \mathcal{D}) \times \pi(w(0), w(1)|\mathcal{D}), \tag{A.9}$$

where

$$w(0) = P_0 + C_1, \qquad w(1) = C_0 + P_1.$$

We have

$$(P_0, D_0, I_0, C_0)|(\mathbf{p}^*, N_0) \sim \text{Multinomial}_{N_0}(p_{10}^*/2, p_{11}, p_{00}, p_{10}^*/2),$$
$$(C_1, D_1, I_1, P_1)|(\mathbf{p}^*, N_1) \sim \text{Multinomial}_{N_1}(p_{10}^*/2, p_{11}, p_{00}, p_{10}^*/2),$$

Algorithm 6 "No benefit" ($\eta_e = 0$) data-augmentation algorithm

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters (μ_0, μ_s, n_0, n_s), desired number of posterior samples T, number of burn-in iterations B, and parameter initialization $(\theta_0^{(0)}, \eta_s^{(0)}) \in (0,1)^2$.

Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $C_1^{(t)}$ conditional on $(\theta_0^{(t-1)}, \eta_s^{(t-1)}, \mathcal{D})$ from the binomial distribution

$$C_1^{(t)} \sim \text{Binomial}\left(y_1, \frac{(1-\theta_0^{(t-1)})\eta_s^{(t-1)}}{\theta_1^{(t-1)}}\right), \tag{A.8}$$

where $\theta_1^{(t-1)} = \theta_0^{(t-1)} + (1-\theta_0^{(t-1)})\eta_s^{(t-1)}$.

(ii) Sample $(\theta_0^{(t)}, \eta_s^{(t)})$ conditional on $(C_1^{(t)}, \mathcal{D})$ from the independent beta (A.6).

Output: Posterior samples after burn-in $\{(\theta_0^{(t)}, \eta_s^{(t)})\}_{t \in \{B+1, \ldots, T\}}$.

and the two distributions are independent. It follows that

$$\begin{aligned}
P_0|(y_0, \mathbf{p}^*) &\sim \text{Binomial}\left(y_0, \frac{p_{10}^*}{p_{10}^* + 2p_{11}}\right), \\
C_0|(y_0, N_0, \mathbf{p}^*) &\sim \text{Binomial}\left(N_0 - y_0, \frac{p_{10}^*}{p_{10}^* + 2p_{00}}\right), \\
C_1|(y_1, \mathbf{p}^*) &\sim \text{Binomial}\left(y_1, \frac{p_{10}^*}{p_{10}^* + 2p_{11}}\right), \\
P_1|(y_1, N_1, \mathbf{p}^*) &\sim \text{Binomial}\left(N_1 - y_1, \frac{p_{10}^*}{p_{10}^* + 2p_{00}}\right),
\end{aligned}$$

independently. Hence, $w(0)$ and $w(1)$ are distributed independently as

$$\begin{aligned}
w(0)|(y_0, y_1, \mathbf{p}^*) &\sim \text{Binomial}\left(y_0 + y_1, \frac{p_{10}^*}{p_{10}^* + 2p_{11}}\right), \\
w(1)|(\mathcal{D}, \mathbf{p}^*) &\sim \text{Binomial}\left(N_0 + N_1 - y_0 - y_1, \frac{p_{10}^*}{p_{10}^* + 2p_{00}}\right),
\end{aligned}$$

Consequently, we have

$$
\begin{aligned}
&\pi(\mathbf{p}^*|w(0),w(1),\mathcal{D}) \\
&\qquad \propto \pi(w(0),w(1),\mathcal{D}|\mathbf{p}^*) \times \pi(\mathbf{p}^*) \\
&\qquad = \pi(w(0),w(1)|\mathcal{D},\mathbf{p}^*) \times \pi(\mathcal{D}|\mathbf{p}^*) \times \pi(\mathbf{p}^*) \\
&\qquad = \pi(w(0)|y_0,y_1,\mathbf{p}^*) \times \pi(w(1)|\mathcal{D},\mathbf{p}^*) \\
&\qquad\qquad \times \pi(\mathcal{D}|\mathbf{p}^*) \times \pi(\mathbf{p}^*) \\
&\qquad = \text{Binomial}\left(w(0); y_0+y_1, \frac{p^*_{10}}{p^*_{10}+2p_{11}}\right) \\
&\qquad\qquad \times \text{Binomial}\left(w(1); N_0+N_1-y_0-y_1, \frac{p^*_{10}}{p^*_{10}+2p_{00}}\right) \\
&\qquad\qquad \times \text{Binomial}(y_0; N_0, p^*_{10}/2+p_{11}) \times \text{Binomial}(y_1; N_1, p^*_{10}/2+p_{11}) \\
&\qquad\qquad \times (p^*_{10})^{\mu_e n_e+\mu_s n_s-1} p_{11}^{(1-\mu_e)n_e-1} p_{00}^{(1-\mu_s)n_s-1} \\
&\qquad \propto (p^*_{10})^{w(0)+w(1)+\mu_e n_e+\mu_s n_s-1} p_{11}^{y_0+y_1-w(0)+(1-\mu_e)n_e-1} p_{00}^{N_0+N_1-y_0-y_1-w(1)+(1-\mu_s)n_s-1}
\end{aligned}
$$

It follows that

$$
\mathbf{p}^*|(w(0),w(1),\mathcal{D}) \sim \text{Dirichlet}(a_{00},a_{10},a_{11}), \tag{A.10}
$$

where

$$
\begin{aligned}
a_{00} &= N_0+N_1-y_0-y_1-w(1)+(1-\mu_s)n_s, \\
a_{10} &= w(0)+w(1)+\mu_e n_e+\mu_s n_s, \\
a_{11} &= y_0+y_1-w(0)+(1-\mu_e)n_e.
\end{aligned}
$$

Consequently, for the mixture weights we have

$$\begin{aligned}
\pi(w(0), w(1)|\mathcal{D}) &= \int \pi(w(0), w(1), \mathbf{p}^*|\mathcal{D}) d\mathbf{p}^* \\
&= \int \pi(w(0), w(1)|\mathbf{p}^*, \mathcal{D}) \pi(\mathbf{p}^*|\mathcal{D}) d\mathbf{p}^* \\
&\propto \binom{y_0 + y_1}{w(0)} \binom{N_0 + N_1 - y_0 - y_1}{w(1)} \\
&\quad \times \int (p_{10}^*/2)^{w(0)+w(1)+\mu_e n_e + \mu_s n_s - 1} p_{11}^{y_0 + y_1 - w(0) + (1-\mu_e) n_e - 1} p_{00}^{N_0 + N_1 - y_0 - y_1 - w(1) + (1-\mu_s) n_s - 1} d\mathbf{p}^* \\
&\propto 2^{-(w(0)+w(1))} \binom{y_0 + y_1}{w(0)} \binom{N_0 + N_1 - y_0 - y_1}{w(1)} \mathrm{B}(a_{00}, a_{10}, a_{11}). \qquad \text{(A.11)}
\end{aligned}$$

Algorithm 7 defines the procedure to sample from the distribution A.9 based on these calculations. Algorithm 8 defines the corresponding data-augmented Gibbs sampler.

A.3 Alternative models and BREASE priors

A.3.1 Other priors for H_0

Recalling that $\theta_0 = p_{10} + p_{11}$ and $\theta_1 = p_{01} + p_{11}$, we see that $\theta_0 = \theta_1$ if and only if $p_{10} = p_{01}$. In this light, we discuss some alternate priors that conform to these constraints. While instantiating H_0 using the beta-binomial model M_0 (4.38) should be preferable in most applications, the prior we discuss here may apply in cases where one has stronger prior information concerning the efficacy and side effects of treatment (η_e, η_s) rather than the baseline risk θ_0 itself.

Aggregated Dirichlet With a Dirichlet$^*(\mu_0, \mu_e, \mu_s; n_0)$ prior on $\mathbf{p}$, we have by the aggregation property of the Dirichlet distribution (Ng et al., 2011)

$$(p_{00}, p_{10} + p_{01}, p_{11}) \sim \mathrm{Dirichlet}((1-\mu_s)n_s, \mu_e n_e + \mu_s n_s, (1-\mu_e)n_e),$$

Algorithm 7 Alternate $H_0 : \theta_0 = \theta_1$ exact posterior sampling algorithm

Input: Data (y_0, y_1, N_0, N_1), hyperparameters (μ_e, μ_s, n_e, n_s), and posterior samples T.

Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $w(1) \in \{0, \ldots, N_0 + N_1 - y_0 - y_1\}$ conditional on (y_0, y_1, N_0, N_1) as

$$\pi(w(1)|y_0, y_1, N_0, N_1) = \sum_{w(0)=0}^{y_0+y_1} \pi(w(0), w(1)|y_0, y_1, N_0, N_1).$$

(ii) Sample $w(0) \in \{0, \ldots, y_0 + y_1\}$ conditional on $(w(1), y_0, y_1, N_0, N_1)$ with probability

$$\pi(w(0)|w(1), y_0, y_1, N_0, N_1) \propto \pi(w(0), w(1)|y_0, y_1, N_0, N_1).$$

(iii) Sample $\mathbf{p}^* = (p_{00}, p^*_{10}, p_{11})$ conditional on $(w(0), w(1), y_0, y_1, N_0, N_1)$ from the Dirichlet distribution (A.10).

(iv) Transform $\mathbf{p}^*$ to obtain samples of $(\theta_0, \theta_1, \eta_e, \eta_s)$ via

$$\theta_0 = p^*_{10}/2 + p_{11} = \theta_1, \qquad \eta_e = \frac{p^*_{10}}{p^*_{10} + 2p_{11}}, \qquad \eta_s = \frac{p^*_{10}}{p^*_{10} + 2p_{00}}.$$

Output: Posterior samples $\{((\mathbf{p}^*)^{(t)}, \theta_0^{(t)}, \theta_1^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})\}_{t \in \{1,\ldots,T\}}$.

Algorithm 8 Alternate $H_0 : \theta_0 = \theta_1$ data-augmentation algorithm

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters (μ_e, μ_s, n_e, n_s), number of posterior samples T, number of burn-in iterations B, and simplex parameter initialization $(\mathbf{p}^*)^{(0)}$.

Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $(w(0)^{(t)}, w(1)^{(t)})$ conditional on $((\mathbf{p}^*)^{(t-1)}, \mathcal{D})$ from the independent binomial

$$w(0)^{(t)} | (y_0, y_1, (\mathbf{p}^*)^{(t-1)}) \sim \text{Binomial}\left(y_0 + y_1, \frac{(p_{10}^*)^{(t-1)}}{(p_{10}^*)^{(t-1)} + 2p_{11}^{(t-1)}}\right),$$
$$w(1)^{(t)} | (\mathcal{D}, (\mathbf{p}^*)^{(t-1)}) \sim \text{Binomial}\left(N_0 + N_1 - y_0 - y_1, \frac{(p_{10}^*)^{(t-1)}}{(p_{10}^*)^{(t-1)} + 2p_{00}^{(t-1)}}\right).$$

(ii) Sample $(\mathbf{p}^*)^{(t)}$ conditional on $(w(0)^{(t)}, w(1)^{(t)}, \mathcal{D})$ from the Dirichlet (A.10).

(iii) Transform $(\mathbf{p}^*)^{(t)}$ to obtain samples $(\theta_0^{(t)}, \theta_1^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})$ via

$$\theta_0 = p_{10}^*/2 + p_{11} = \theta_1, \qquad \eta_e = \frac{p_{10}^*}{p_{10}^* + 2p_{11}}, \qquad \eta_s = \frac{p_{10}^*}{p_{10}^* + 2p_{00}}.$$

Output: Posterior samples after burn-in $\{((\mathbf{p}^*)^{(t)}, \theta_0^{(t)}, \theta_1^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})\}_{t \in \{B+1, \ldots, T\}}$.

where $n_e = \mu_0 n_0$ and $n_s = (1-\mu_0)n_0$. Assuming H_0 holds, and defining $p_{10}^* = p_{10} + p_{01} = 2p_{10}$, we obtain the Dirichlet prior density on the aggregated cell probabilities

$$\pi(p_{00}, p_{10}^*) = \text{B}((1-\mu_s)n_s, \mu_e n_e + \mu_s n_s, (1-\mu_e)n_e)^{-1} p_{00}^{(1-\mu_s)n_s - 1} (p_{10}^*)^{\mu_e n_e + \mu_s n_s - 1} p_{11}^{(1-\mu_e)n_e - 1},$$

where $p_{11} = 1 - p_{00} - p_{10}^*$ and $\text{B}(a_{00}, a_{10}, a_{11})$ is the multivariate beta function:

$$\text{B}(a_{00}, a_{10}, a_{11}) = \frac{\Gamma(a_{00})\Gamma(a_{10})\Gamma(a_{11})}{\Gamma(a_{00} + a_{10} + a_{11})}.$$

This prior allows for exact posterior sampling and marginal likelihood calculation in cases where we may have stronger prior information concerning the efficacy and side effects of treatment (η_e, η_s) than the baseline risk θ_0. Indeed, note that the prior is fully specified

by the hyperparameters (μ_e, μ_s, n_e, n_s). Recalling that the Dirichlet* prior is obtained from the generalized Dirichlet by setting $n_e = \mu_0 n_0$ and $n_s = (1-\mu_0)n_0$, we see that this prior assumes that we have as much prior knowledge on θ_0 as we do on (η_e, η_s).

With this parametrization, the likelihood under H_0 is given by

$$L(\mathcal{D}|p) = \binom{N_0}{y_0}\binom{N_1}{y_1}(p_{10}^*/2 + p_{11})^{y_0+y_1}(p_{00} + p_{10}^*/2)^{N_0+N_1-y_0-y_1}.$$

The posterior is then

$$\begin{aligned}\pi(p_{00}, p_{10}^*|\mathcal{D}) \propto \binom{N_0}{y_0}\binom{N_1}{y_1}&\mathrm{B}((1-\mu_s)n_s, \mu_e n_e + \mu_s n_s, (1-\mu_e)n_e)^{-1} \\ &\times (p_{10}^*/2 + p_{11})^{y_0+y_1}(p_{00} + p_{10}^*/2)^{N_0+N_1-y_0-y_1} \\ &\times (p_{10}^*)^{\mu_e n_e + \mu_s n_s - 1} p_{11}^{(1-\mu_e)n_e - 1} p_{00}^{(1-\mu_s)n_s - 1}.\end{aligned}$$

From here we can apply the binomial theorem twice to quickly see that the posterior is a mixture of Dirichlet densities on the probability vector $\mathbf{p}^* = (p_{00}, p_{10}^*, p_{11})$. This yields the marginal likelihood formula

$$\begin{aligned}L(\mathcal{D}) = &\binom{N_0}{y_0}\binom{N_1}{y_1}\mathrm{B}((1-\mu_s)n_s, \mu_e n_e + \mu_s n_s, (1-\mu_e)n_e)^{-1} \\ &\times \sum_{j=0}^{y_0+y_1}\sum_{k=0}^{N_0+N_1-y_0-y_1} 2^{-(j+k)}\binom{y_0+y_1}{j}\binom{N_0+N_1-y_0-y_1}{k}\mathrm{B}(a_{00}(j,k), a_{10}(j,k), a_{11}(j,k)),\end{aligned}$$

where we define

$$\begin{aligned}a_{00}(j,k) &= N_0 + N_1 - y_0 - y_1 + (1-\mu_s)n_s - k, \\ a_{10}(j,k) &= j + k + \mu_e n_e + \mu_s n_s, \\ a_{11}(j,k) &= y_0 + y_1 + (1-\mu_e)n_e - j.\end{aligned}$$

In Section A.2.3, we derive an algorithm for exact posterior sampling using the aggregated Dirichlet prior on $(p_{00}, p_{10}^*, p_{11})$.

A.3.2 Other priors for H_- and H_+

Another approach for specifying models for H_- and H_+, which is both natural and computationally convenient, is to impose a monotonicity assumption on M_1, and set $\eta_s = 0$ or $\eta_e = 0$ respectively. This results in the following models,

$$M_-^{'}: \ (\theta_0, \eta_e) \sim \text{Beta}^*(\mu_0, n_0) \times \text{Beta}^*(\mu_e, n_e), \qquad \theta_1 = (1 - \eta_e)\theta_0 \tag{A.12}$$

$$M_+^{'}: \ (\theta_0, \eta_s) \sim \text{Beta}^*(\mu_0, n_0) \times \text{Beta}^*(\mu_s, n_s), \qquad \theta_1 = \theta_0 + \eta_s(1 - \theta_0), \tag{A.13}$$

with marginal likelihoods given by

$$\begin{aligned} L_-^{'}(\mathcal{D}) = &\binom{N_0}{y_0}\binom{N_1}{y_1}\sum_{k=0}^{N_1-y_1}\binom{N_1-y_1}{k} \\ &\times \frac{\text{B}(y_0 + y_1 + k + \mu_0 n_0, N - (y_0 + y_1 + k) + (1 - \mu_0)n_0)}{\text{B}(\mu_0 n_0, (1 - \mu_0)n_0)} \\ &\times \frac{\text{B}(k + \mu_e n_e, y_1 + (1 - \mu_e)n_e)}{\text{B}(\mu_e n_e, (1 - \mu_e)n_e)}, \end{aligned}$$

and

$$\begin{aligned} L_+^{'}(\mathcal{D}) = &\binom{N_0}{y_0}\binom{N_1}{y_1}\sum_{j=0}^{y_1}\binom{y_1}{j} \\ &\times \frac{\text{B}(y_0 + j + \mu_0 n_0, N - (y_0 + j) + (1 - \mu_0)n_0)}{\text{B}(\mu_0 n_0, (1 - \mu_0)n_0)} \\ &\times \frac{\text{B}(y_1 - j + \mu_s n_s, N_1 - y_1 + (1 - \mu_s)n_s)}{\text{B}(\mu_s n_s, (1 - \mu_s)n_s)}. \end{aligned}$$

Here we interpret the constraint $\eta_s = 0$ (or $\eta_e = 0$) simply as a causally principled way to derive a prior compatible with the desired constraint $H_- : \theta_1 < \theta_0$ (or $H_+ : \theta_1 > \theta_0$), and not as testing the former constraint in lieu of the latter.[1] One interesting characteristic of

[1] In general, the data cannot differentiate the stronger constraint, such as $\eta_s = 0$ (no one is hurt by the treatment), from the weaker constraint $\theta_1 < \theta_0$ (the treatment is beneficial on average), since the likelihood depends only on θ_1 and θ_0. Thus, in this case, differences in using M_- or $M_-^{'}$ amounts to differences only in the induced priors satisfying the same testable constraint $\theta_1 < \theta_0$, such as one placing more (or less) mass on smaller (or larger) effects than the other.

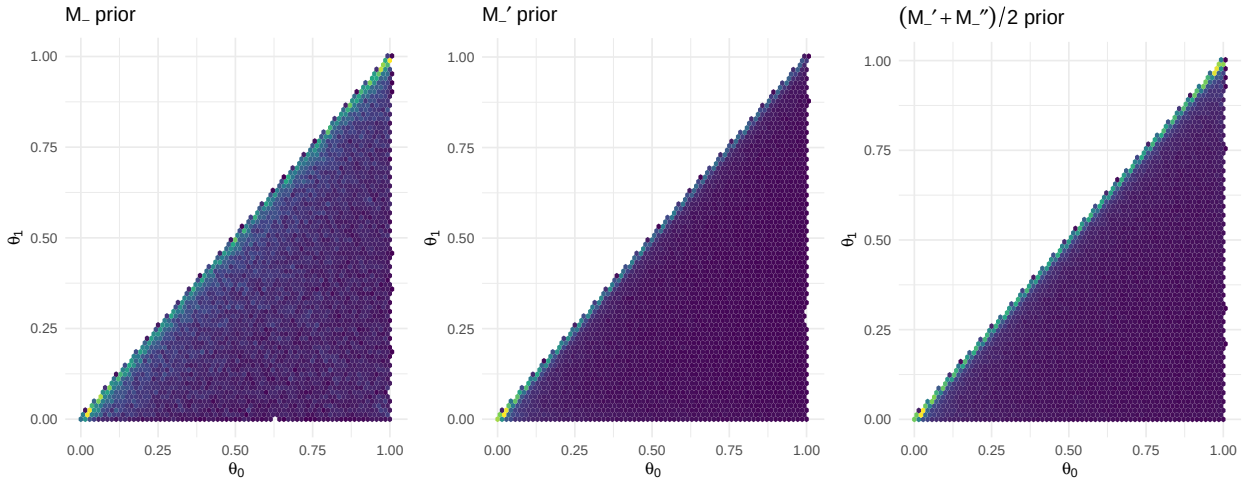


Figure A.52: Left: heatmap of joint prior on (θ_0, θ_1) implied by the M_- prior (4.39) with $\mu_0 = 1/2, \mu_e = \mu_s = 0.3, n_0 = 2, n_e = n_s = 1$. Center: prior on (θ_0, θ_1) under $M_-^{'}$ with the same values of (μ_0, μ_e, n_0, n_e). Right: prior on (θ_0, θ_1) under the mixture model $(M_-^{'} + M_-^{''})/2$ with $\mu_1 = 1/2, \mu_s^{'} = 0.3, n_1 = 2, n_s^{'} = 1$ and the same values of (μ_0, μ_e, n_0, n_e).

models $M_-^{'}$ and $M_+^{'}$ is that they do not put θ_0 and θ_1 on equal footing, even when choosing beta priors compatible with the BREASE$(1/2, \mu, \mu; 2, 1, 1)$ distribution, which places flat marginals on θ_0 and θ_1. This is usually desirable, e.g., when the control condition indeed denotes a well understood baseline, such as a standard of care. Symmetry of θ_0 and θ_1, however, can also be easily restored by switching the roles of the "treatment" and "control" conditions, as discussed in Appendix A.3.2. Algorithms to sample exactly from the posterior under $M_-^{'}$ and $M_+^{'}$ are provided in Appendix A.2.

Returning to the model $M_-^{'}$ $(A.12)$, some natural values for the prior hyperparameters are

$$\mu_0 = \mu_e = 1/2, \qquad n_0 = n_e = 2,$$

which define a flat Uniform$(0,1)^2$ prior on (θ_0, η_e). The resulting conditional prior on θ_1 is

$$\theta_1|\theta_0 \sim \text{Uniform}(0, \theta_0),$$

which presents an intuitive representation of the hypothesis $H_- : \theta_1 < \theta_0$. Note, however, that this specification of the model handles θ_0 as the baseline quantity. We can also go in the other direction, specifying priors on θ_1 and the "side effects of placebo" η_s' and defining

$$\theta_0 = \theta_1 + (1 - \theta_1)\eta_s',$$

which also instantiates $H_- : \theta_1 < \theta_0$. We denote by M_-'' the model

$$(\theta_1, \eta_s') \sim \text{Beta}^*(\mu_1, n_1) \times \text{Beta}^*(\mu_s', n_s'),$$
$$\theta_0 = \theta_1 + (1 - \theta_1)\eta_s'.$$

This asymmetry in our handling of θ_0 and θ_1 is reflected in the joint priors of (θ_0, θ_1) under M_-' and M_-''. As the central panel of Figure A.52 exhibits, the M_-' joint prior tends to favor small proportions (whereas M_-'', not plotted, favors large proportions). On the other hand, sampling $(\theta_0, \eta_e, \eta_s)$ from the BREASE prior truncated to the set $\{(\theta_0, \eta_e, \eta_s) : \theta_1 < \theta_0\}$ (i.e., the M_- prior (4.39)) yields a symmetric joint density on (θ_0, θ_1) (left panel of Figure A.52). To assuage this asymmetry, we can put θ_0 and θ_1 on equal footing when testing the one-sided hypothesis H_- (and, similarly, H_+) by using a prior that averages those under M_-' and M_-'', as in the right panel of Figure A.52. In practice, we can decompose H_- into the submodels M_-' and M_-'' and report the marginal likelihood of H_- as the average of the submodel marginal likelihoods. As the marginal likelihood under M_-'' is also available analytically, this procedure comes with negligible added computational cost.

A.3.3 An empirical Bayes prior

As η_e and η_s are counterfactual probabilities, they are not generally point-identified from data. However, since θ_0 and θ_1 are identifiable, we can derive robust bounds on their range

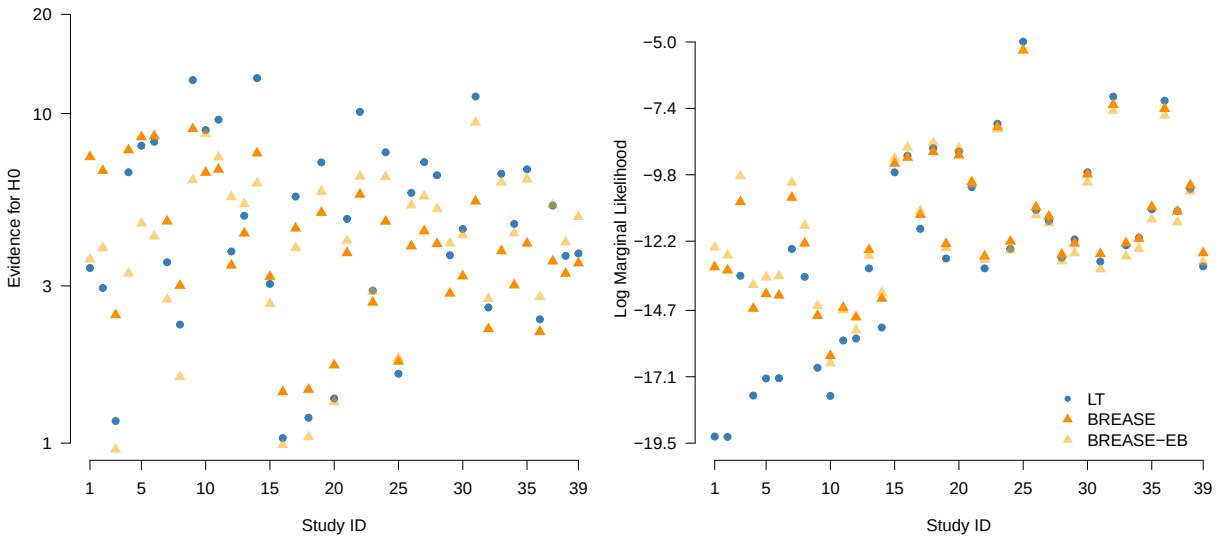


Figure A.53: Comparison of Bayes factors (BF_{01}) and log marginal likelihoods under model M_1 (4.37) of the default LT, default BREASE, and empirical Bayes BREASE priors across the 39 *NEJM* studies.

of possible values based on the observed data (Tian & Pearl, 2000). Equation (4.8) implies the following algebraic constraints on η_e and η_s:

$$\max\left\{0, \frac{\theta_0 - \theta_1}{\theta_0}\right\} \leq \eta_e \leq \min\left\{1, \frac{1-\theta_1}{\theta_0}\right\}, \tag{A.14}$$

$$\max\left\{0, \frac{\theta_1 - \theta_0}{1-\theta_0}\right\} \leq \eta_s \leq \min\left\{1, \frac{\theta_1}{1-\theta_0}\right\}. \tag{A.15}$$

The inequalities (A.14) and (A.15) define the (marginal) partially identified regions of η_e and η_s, respectively. Denote these intervals by $I_e(\theta_0, \theta_1) = [\ell_e(\theta_0, \theta_1), u_e(\theta_0, \theta_1)]$ and $I_s(\theta_0, \theta_1) = [\ell_s(\theta_0, \theta_1), u_s(\theta_0, \theta_1)]$. In the limit of infinite data, the posterior mass of η_e and η_s will concentrate within $I_e(\theta_0^*, \theta_1^*)$ and $I_s(\theta_0^*, \theta_1^*)$, respectively, assuming θ_0^*, θ_1^* are the true values.

When conducting a Bayesian hypothesis test, a main concern is the sensitivity of Bayes factors to the prior. As demonstrated in Section 3.3, a prior that places unreasonable as-

sumptions on the treatment effects can lead to questionable conclusions. In this light, it may be desired to take a data-driven approach to prior specification that concentrates prior mass near the partially identified intervals of η_e and η_s. For example, we can set the prior means μ_e and μ_s to equal their midpoints:

$$\hat{\mu}_e = \frac{1}{2}\left(\ell_e(\hat{\theta}_0, \hat{\theta}_1) + u_e(\hat{\theta}_0, \hat{\theta}_1)\right),$$
$$\hat{\mu}_s = \frac{1}{2}\left(\ell_s(\hat{\theta}_0, \hat{\theta}_1) + u_s(\hat{\theta}_0, \hat{\theta}_1)\right),$$

where we use point estimates of the population proportions:

$$\hat{\theta}_0 = \frac{y_0 + 1}{N_0 + 2}, \qquad \hat{\theta}_1 = \frac{y_1 + 1}{N_1 + 2}.$$

As $\hat{\theta}_0$ shrinks the sample proportion toward 1/2, it avoids division by zero in (A.14) and (A.15). Hence, we might consider priors of the form BREASE$(1/2, \hat{\mu}_e, \hat{\mu}_s; 2, n, n)$ with $n \geq 0$. As this prior is estimated from the observed data, it can be thought of as an empirical Bayes approach (Robbins, 1992). As such, we denote it by BREASE-EB(n).

Note that when $n = 1$ and $\hat{\theta}_0 = \hat{\theta}_1 = 1/2$ (e.g., in the absence of data or when the sample proportions are 1/2), we obtain a vague Jeffreys marginal prior Beta$(1/2, 1/2)$ on η_e and η_s. The choice of prior sample size $n = 1$ yields something resembling a unit information prior (Kass & Wasserman, 1995), wherein the prior mean is estimated from data and its spread is chosen so that the information content of the prior matches that of a single observation.

Figure A.53 compares Bayes factors (BF_{01}) and log marginal likelihoods under model M_1 (4.37) of the default LT$(0, 0; 1, 1)$, BREASE$(1/2, 0.3, 0.3; 2, 1, 1)$, and BREASE-EB(1) priors across the 39 *NEJM* studies reporting null results. The BREASE and BREASE-EB priors tend to provide the most equivocal Bayes factors on average, with mean BF_{01} equal to 4.41, 4.42, and 5.38 for the BREASE-EB, BREASE, and LT priors, respectively. However, BREASE-EB Bayes factors tend to be closer to those of the LT approach than the default BREASE prior, with mean absolute percentage differences from the LT BF_{01} of 19% for the former and 32% for the latter.

Comparing log marginal likelihoods, which quantify the predictive performance of a model, we see that the BREASE-EB and default BREASE priors perform similarly, and generally better than the default LT prior, although the default BREASE performs slightly better overall. Indeed, the default BREASE log marginal likelihood exceeds the LT in 74% of the studies compared to 59% for the BREASE-EB prior. Furthermore, the default BREASE outperforms BREASE-EB in 62% of the studies.