







Methods

Forecasting COVID-19 and Analyzing the Effect of Government Interventions

Michael Lingzhi Li,^a Hamza Tazi Bouardi,^a Omar Skali Lami,^a Thomas A. Trikalinos,^b Nikolaos Trichakis,^c Dimitris Bertsimas^{c,*}

^aOperations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; ^bCenter for Evidence Synthesis in Health, Brown University, Providence, Rhode Island 02912; ^cSloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

*Corresponding author

Contact: mlli@mit.edu,  <https://orcid.org/0000-0002-2456-4834> (MLL); htazi@mit.edu,  <https://orcid.org/0000-0002-7871-325X> (HTB); oskali@mit.edu,  <https://orcid.org/0000-0002-8208-3035> (OSL); thomas_trikalinos@brown.edu,  <https://orcid.org/0000-0002-3990-1848> (TAT); ntrichakis@mit.edu,  <https://orcid.org/0000-0002-8324-9148> (NT); dbertsim@mit.edu,  <https://orcid.org/0000-0002-1985-1003> (DB)

Received: July 8, 2020

Revised: February 21, 2021; December 1, 2021

Accepted: April 25, 2022

Published Online in *Articles in Advance*:
June 10, 2022

Area of Review: Machine Learning and Data
Science.

<https://doi.org/10.1287/opre.2022.2306>

Copyright: © 2022 INFORMS

Abstract. We developed DELPHI, a novel epidemiological model for predicting detected cases and deaths in the prevaccination era of the COVID-19 pandemic. The model allows for underdetection of infections and effects of government interventions. We have applied DELPHI across more than 200 geographical areas since early April 2020 and recorded 6% and 11% two-week, out-of-sample median mean absolute percentage error on predicting cases and deaths, respectively. DELPHI compares favorably with other top COVID-19 epidemiological models and predicted in 2020 the large-scale epidemics in many areas, including the United States, United Kingdom, and Russia, months in advance. We further illustrate two downstream applications of DELPHI, enabled by the model’s flexible parametric formulation of the effect of government interventions. First, we quantify the impact of government interventions on the pandemic’s spread. We predict, that in the absence of any interventions, more than 14 million individuals would have perished by May 17, 2020, whereas 280,000 deaths could have been avoided if interventions around the world had started one week earlier. Furthermore, we find that mass gathering restrictions and school closings were associated with the largest average reductions in infection rates at $29.9 \pm 6.9\%$ and $17.3 \pm 6.7\%$, respectively. The most stringent policy, stay at home, was associated with an average reduction in infection rate by $74.4 \pm 3.7\%$ from baseline across countries that implemented it. In the second application, we demonstrate how DELPHI can predict future COVID-19 incidence under alternative governmental policies and discuss how Janssen Pharmaceuticals used such analyses to select the locations of its Phase III trial for its leading single-dose vaccine candidate Ad26.Cov2.S.

Supplemental Material: The e-companion is available at <https://doi.org/10.1287/opre.2022.2306>.

Keywords: epidemiology • compartmental modeling • infectious diseases • partial identifiability

1. Introduction

The ongoing COVID-19 pandemic is the deadliest in recent history. As of June 1st, 2022, there were more than 530 million confirmed cases of COVID-19 and 6.3 million deaths. In late March 2020, we developed DELPHI, a new epidemiological model that aims to predict the pandemic’s evolution. DELPHI extends a classical SEIR model (Kermack and McKendrick 1927) to include additional outcomes, such as deaths, account for underdetection of infections, and estimate the effect of changing government interventions. Since its inception, DELPHI has been one of the top four models consistently incorporated into the U.S. Centers for Disease Control and Prevention’s (CDC) core ensemble forecast (Dean et al.

2020) and has been utilized by various health and federal agencies, including the Federal Reserve, for pandemic planning. DELPHI was used by Janssen Pharmaceuticals to select the locations of its multicenter Phase III trial for its single-dose vaccine candidate Ad26.Cov2.S and by Hartford Healthcare, a major hospital system in the United States, to plan intensive care unit capacity.

A key strength of DELPHI is its explicit, flexible, and parametric modeling of government interventions. During the COVID-19 pandemic, governments around the world enacted wide-ranging nonpharmaceutical interventions (NPIs), including social distancing, school closures, and lockdowns, at different stages of their local

epidemics. This variation helps DELPHI to identify the effects of different interventions, which, in turn, allows for predictions of alternative scenarios and can inform policymaking.

We demonstrate two applications of DELPHI that leverage our modeling of government interventions. First, through parameter calibration to time series of observed cases and deaths in various countries, DELPHI can estimate the effect of different NPIs accounting for country-specific baseline infection and case fatality rates. In Section 4.1.1, we estimate that school closings and mass gathering restrictions were among the most effective measures in reducing the rate of infection during the early stages of the pandemic. Although these policies incur a significant social burden, they can be effective in controlling the extreme growth in infections when other preventative measures (e.g., masks) are not widely implementable and while treatment options are being evaluated and developed. Had these restrictions been implemented just one week earlier, most — up to 90% — of the deaths in the early stages of the pandemic could have been avoided.

Another major application of DELPHI is the assessment of alternative scenarios to inform policymaking. By utilizing the estimated effect of different NPIs, we can create a scenario analysis toolkit to simulate the pandemic forward under different government policies. In Section 4.2, we illustrate Janssen Pharmaceuticals' (a Johnson & Johnson company) use of this scenario analysis toolkit in mid-to-late 2020 to identify countries with predicted high COVID-19 incidence as candidate sites for their multicenter Phase III randomized trial of their leading vaccine candidate Ad26.Cov2.S.

DELPHI has been applied to 167 geographic areas (countries/provinces/states) worldwide as of end of April 2020 and more than 215 as of end of September 2020, covering all six populated continents. Its results have also been available since early April 2020 on www.covidanalytics.io. In this paper, we document the calibration, quantitative results, and insights obtained from the DELPHI model during the prevaccination era of the COVID-19 epidemic and illustrate two key applications.

1.1. Literature

Many epidemiological models were developed to describe the evolution of the COVID-19 epidemic. Most are mechanistic and represent some variation of the classical Susceptible-Exposed-Infectious-Recovered (SEIR) compartmental model (Kermack and McKendrick 1927), which partitions a population into mutually exclusive and exhaustive compartments and describes infection dynamics with differential equations. Some of these models are marginal in that they do not have different compartments for different age, sex, or occupation strata (Gu 2020), whereas others account for population

substructure (PSI-DRAFT 2020). A subset of models parameterize the force of the infection as a function of predictors such as proxies of behavior (e.g., cell phone-derived mobility data, credit card spending data) or governmental policies (Chinazzi et al. 2020, Woody et al. 2020). Nonmechanistic models use machine-learning (Rodriguez et al. 2020) or statistical time-series modeling to forecast the evolution of outcomes (Mehrotra and Ivan 2020). For a comprehensive review of COVID-19 models, see Dean et al. (2020).

DELPHI is a mechanistic compartmental model. For each country or state, it describes government policies as composites of elemental NPIs, including lockdown, school closures, gathering size restrictions, and restrictions on nonessential businesses. DELPHI parameterizes the net impact of government policies on the activity of the epidemic as a nonlinear function of the effects of the elemental NPIs and estimates the latter by fitting them to time series of COVID-19 cases and deaths. DELPHI also accounts for the improvement in the management of COVID-19 patients over time by modeling the mortality of the disease as a function of time and allows for different dynamics for those who recover versus those who die from the disease. It is one of the models most consistently included in the CDC ensemble forecast (top 4 out of 30 submitted models) (Ray et al. 2020), and its favorable performance is demonstrated in Section 3.2.

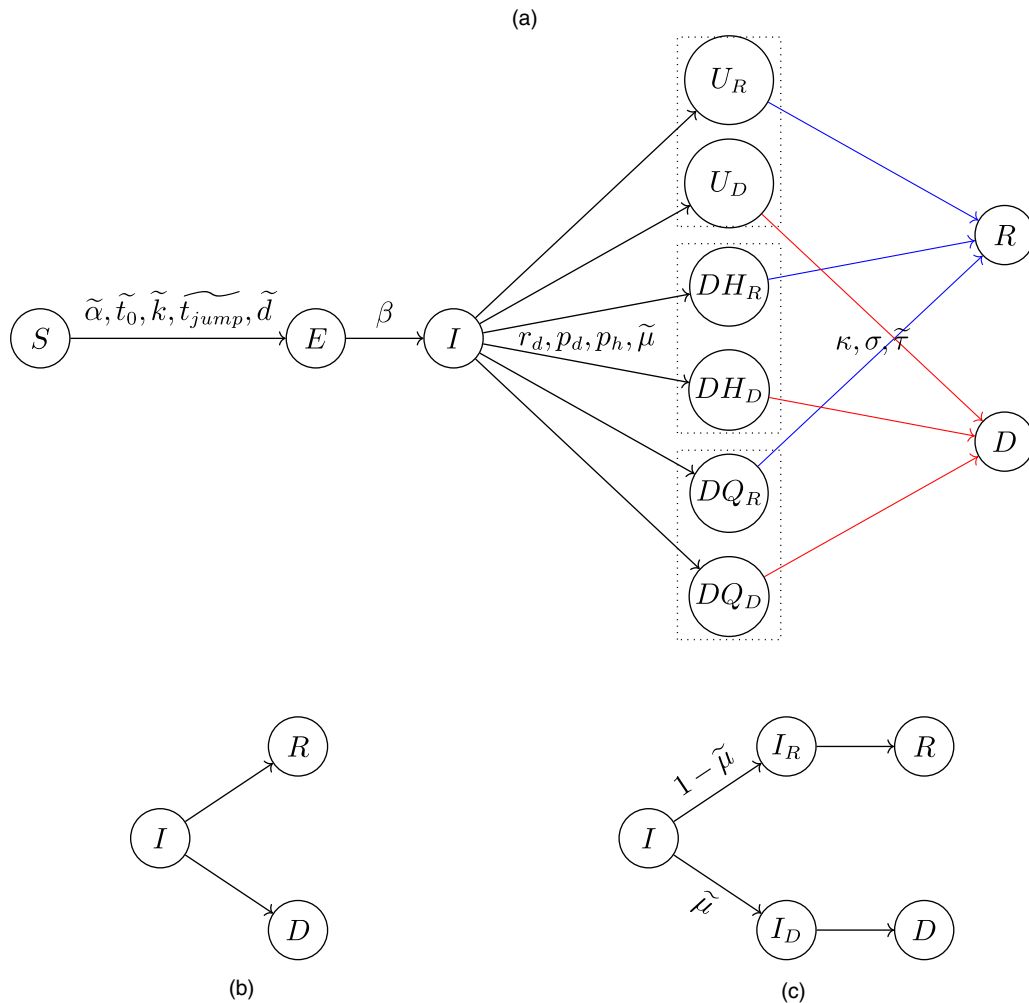
We believe that the flexibility of the aforementioned modeling choices largely belies the successful projections of the epidemic's trajectory in more than 200 countries on all six populated continents. Furthermore, by estimating the effects of elemental NPIs, we can inform policymaking (Section 4.1.1), estimate the impact of delays in deploying early interventions (Section 4.1.2), and inform the design of clinical trials (Section 4.2).

2. The DELPHI Model

DELPHI is a closed system in that it does not include demography (births, non-COVID-19 deaths) or migration. It partitions the population in 11 mutually exclusive and exhaustive compartments (Figure 1(a)):

- **Susceptible (S):** People who are susceptible to SARS-CoV-2 infection.
- **Exposed (E):** Early-infected persons who are not yet contagious and are in the incubation period.
- **Infected (I):** People currently infected and contagious.
- **Undetected (U_R and U_D):** Infected people who are not detected (and are not counted among the known cases) because they did not get tested. It is assumed that these compartments do not contribute to the infection process because the corresponding persons develop symptoms and are then self-quarantined. Some will die

Figure 1. The DELPHI Model



Note: (a) Flow Diagram of DELPHI; (b) standard modeling for recoveries and deaths; (c) updated modeling for recoveries and deaths

(U_D) with marginal probability of death $\tilde{\mu}(t)$, and the rest will recover.

- **Detected, Hospitalized (DH_R and DH_D):** People who are infected, confirmed, and have severe enough disease to be hospitalized (and effectively quarantined and not contributing to the infectious process). Some will die (DH_D) with marginal probability of death $\tilde{\mu}(t)$, and the rest will recover (DH_R).

- **Detected, Quarantined (DQ_R and DQ_D):** People who are infected, detected through testing, and home-quarantined so that they do not infect others. Some of these people will die (DQ_D) with marginal probability of death $\tilde{\mu}(t)$, and the rest will recover (DQ_R).

- **Recovered (R):** People who have recovered from the disease and are assumed to have acquired natural immunity. This immunity does not wane for the time horizon of the model.

- **Deceased (D):** People who have died from COVID-19.

In a given area (state/country), the full mathematical formulation of the model is as follows:

$$\frac{dS}{dt} = -\tilde{\alpha}\gamma(t)S(t)I(t)$$

$$\frac{dE}{dt} = \tilde{\alpha}\gamma(t)S(t)I(t) - \beta E(t)$$

$$\frac{dI}{dt} = \beta E(t) - r_d I(t)$$

$$\frac{dU_R}{dt} = r_d(1 - \tilde{\mu}(t))(1 - p_d)I(t) - \sigma U_R(t)$$

$$\frac{dDH_R}{dt} = r_d(1 - \tilde{\mu}(t))p_d p_h I(t) - \kappa DH_R(t)$$

$$\frac{dDQ_R}{dt} = r_d(1 - \tilde{\mu}(t))p_d(1 - p_h)I(t) - \sigma DQ_R(t)$$

$$\begin{aligned}\frac{dU_D}{dt} &= r_d \tilde{\mu}(t)(1 - p_d)I(t) - \tilde{\tau}U_D(t) \\ \frac{dDH_D}{dt} &= r_d \tilde{\mu}(t)p_d p_h I(t) - \tilde{\tau}DH_D(t) \\ \frac{dDQ_D}{dt} &= r_d \tilde{\mu}(t)p_d(1 - p_h)I(t) - \tilde{\tau}DQ_D(t) \\ \frac{dTH}{dt} &= r_d p_d p_h I(t) \\ \frac{dDD}{dt} &= \tilde{\tau}(DH_D(t) + DQ_D(t)) \\ \frac{dDT}{dt} &= r_d p_d I(t) \\ \frac{dR}{dt} &= \sigma(U_R(t) + DQ_R(t)) + \kappa DH_R(t) \\ \frac{dD}{dt} &= \tilde{\tau}(U_D(t) + DQ_D(t) + DH_D(t)).\end{aligned}$$

At the beginning of the epidemic, the system above starts at $t = 0$ with the initial condition $S(0) = N - \tilde{k}_1 - \tilde{k}_2$, $E(0) = \tilde{k}_1$, $I(0) = \tilde{k}_2$, and all other compartments starting at 0, with \tilde{k}_1, \tilde{k}_2 country-specific fitted parameters. As described later in the current section, the model is restarted after an epidemic wave has been observed. For each wave, we reset $t \rightarrow 0$ and use as a starting condition the distribution of the compartments that was reached in the previous period for all compartments except S , E , and I . The starting values of the latter are refit using \tilde{k}_1 and \tilde{k}_2 :

$$S(0) = N - \tilde{k}_1 - \tilde{k}_2 - R(0) - \sum_{i \in \{R, D\}} DH_i(0) + DQ_i(0) + U_i(0), \quad E(0) = \tilde{k}_1, \quad I(0) = \tilde{k}_2.$$

In Figure 1(a), nodes represent compartments and arrows the allowable transitions. The variables that govern the dynamics are listed in proximity to the associated transitions. Variables with a tilde ($\tilde{}$) are fitted to the time series of known cases and deaths in each area (country/state/province). The remaining are fixed to global values informed from a literature review of 174 papers, which was current when DELPHI was developed (Bertsimas et al. 2020):

- $\tilde{\alpha}$ is the baseline infection rate.
- $\gamma(t)$ measures the effect of government response and is defined as

$$\gamma(t) = 1 + \frac{2}{\pi} \arctan\left(\frac{-(t - \tilde{t}_0)}{\tilde{k}}\right) + \tilde{c} \exp\left(-\frac{(t - \tilde{t}_{\text{jump}})^2}{2\tilde{d}^2}\right), \quad (1)$$

where the parameters \tilde{t}_0 and \tilde{k} capture, respectively, the timing and the strength of the response. This function is refit to data when the model is restarted (e.g., with the emergence of a new wave, as described later). The exponential term intends to reflect a resurgence in infections due to relaxation of governmental policy and societal response; \tilde{c} controls the magnitude of the resurgence, \tilde{t}_{jump} the time when the resurgence peaks, and \tilde{d} the duration of the resurgence phase. The effective infection rate in the model is $\tilde{\alpha}\gamma(t)$, which is time dependent. The exponential resurgence term was added to the model in late June 2020, when we observed the first large resurgence in the pandemic (before July the model assumed $\tilde{c} = 0$). The $\frac{2}{\pi}$ constant normalizes the arctan function. For example, before July 2020, during the first wave, $\tilde{c} = 0$ (no resurgence), $\gamma(t)$ had range $[0, 2]$, and $\gamma(t) = 1$ when $t = t_0$.

- r_d is the detection rate. This equals to $\frac{\log 2}{T_d}$, where T_d is the median time to detection (fixed to be 2 days); see Wang et al. (2020).

- β is the rate of infection leaving incubation phase. This equals to $\frac{\log 2}{T_\beta}$, where T_β is the median time to leave incubation (fixed at 5 days), see Lauer et al. (2020).

- σ is the rate of recovery of nonhospitalized patients. This equals to $\frac{\log 2}{T_\sigma}$, where T_σ is the median time to recovery of nonhospitalized patients (fixed at 10 days); see Hu et al. (2020), Kluytmans et al. (2020).

- κ is the rate of recovery under hospitalization. This equals to $\frac{\log 2}{T_\kappa}$, where T_κ is the median time to recovery under hospitalization (fixed at 15 days); see Grein et al. (2020) and Liu et al. (2020b).

- $\tilde{\tau}$ is the death rate, the reciprocal of the average time it takes for patients to move from the U_D, DH_D , and DQ_D to the death compartment D . This captures the speed at which patients die.

- $\tilde{\mu}(t)$ is the time-varying marginal probability of death among all infected patients, also known as the infection fatality rate. This function is refit to data when the model is restarted. It is parameterized as a monotone function of time

$$\tilde{\mu}(t) = (\tilde{\mu}_0 - \mu_{\min}) \left(1 + \frac{2}{\pi} \arctan(-\tilde{r}_m t)\right) + \mu_{\min},$$

where $\tilde{\mu}_0$ is the initial probability of death (at the time the model is (re)started), μ_{\min} is its minimum, and \tilde{r}_m is a daily decay rate for mortality. In most areas, mortality is monotonically decreasing ($\tilde{r}_m > 0$), reflecting improvements in patient management. A negative decay rate would correspond to a worsening of the probability of death over time, as was observed in some areas when the capacity of the local health system was exceeded. In the version of the model before

June 2020, we assumed that $\tilde{r}_m = 0$ because the mortality rate was relatively constant in the early pandemic when optimal patient management was not universally followed:

- p_d is the probability that a contagious person will be detected. It is fixed at 20% based on various early estimations of the detection probability in countries with earlier outbreaks (Krantz and Rao 2020, Niehus et al. 2020, Wang et al. 2020).

- p_h is the probability that a detected case will be hospitalized and is set to 15%; see Arons et al. (2020) and Xu et al. (2020).

DELPHI is fit separately in each area (country/state/province, as applicable) and over successive “training windows” that begin when a new wave has lasted for at least one month. For each area and training window, we fit 13 parameters from the list above (\tilde{k}_1, \tilde{k}_2 for the initial condition and $\tilde{\alpha}, \tilde{\mu}, \tilde{\tau}, \tilde{t}_0, \tilde{k}, \tilde{c}, \tilde{t}_{\text{jump}}, \tilde{d}, \tilde{\mu}_0, \tilde{r}_m$ from the list above) by minimizing a weighted mean squared error (MSE) loss. Let $DT(t)$ and $DD(t)$ denote the number of reported total detected cases and detected deaths, respectively, on day t . Then, the loss function for a training period of T days is defined as

$$\sum_{t=1}^T \frac{t^2}{T^2} \cdot (\widehat{DT}(t) - DT(t))^2 + \lambda^2 \cdot \sum_{t=1}^T \frac{t^2}{T^2} \cdot (\widehat{DD}(t) - DD(t))^2,$$

where $\widehat{DT}(t)$ and $\widehat{DD}(t)$ are predicted detected cases and deaths, respectively. The factor $\frac{t^2}{T^2}$ gives more prominence to more recent data, because recent errors are more likely to propagate into future errors. The lambda factor $\lambda = \min\left\{\frac{DT(T)}{3 \cdot DD(T)}, 10\right\}$ balances the fitting between detected cases and deaths; this rescaling coefficient was obtained experimentally through cross-validation.

The training windows are dynamically updated, with the goal that each should cover a period in which the enacted measures are not strengthened. Manually tracking policy changes to trigger a retraining proven impractical, given that DELPHI was applied to a large number of areas. Therefore, retraining the model was triggered once a new wave that has lasted for at least one month was detected by tracking the data. In that case, the start date of the training period was set as to

be the beginning of that wave. The rationale behind this heuristic is that, usually, stricter measures are enacted once a new wave is evident and on the rise. As an example, for the United States, in July 2020 retraining was triggered, with the new training period starting from June 2020, and similarly in November 2020, with a start from October 2020. We specifically exclude historical data starting before an area recorded more than 100 cases for numerical stability and to exclude sporadic outbreaks before the actual epidemic.

To optimize over the highly nonconvex search space, we utilize both a local truncated Newton algorithm (TNC) (Nocedal and Wright 2006) and a global optimization method of dual annealing (DA) (Xiang et al. 1997). TNC is utilized to produce forecasts on a daily basis, whereas DA, being more computationally expensive, is performed on a weekly basis to shift and readjust the parameters more significantly if the underlying mechanics have changed (e.g., in the case of a new wave of cases). Parameters are fit by using bounds of plus/minus $\pm 20\%$ deviation around the latest value for TNC and bounds of $\pm 50\%$ deviation for DA. When we first trained the model, we used parameter ranges that were obtained from initial estimates derived in studies in South Korea and China, $\pm 20\%$.

We would illustrate this with a specific example focused on the region of Georgia. Table 1 shows the parameters and their bounds for the update on October 23, 2020, when we utilized both TNC and DA. We only show a selection of parameters to improve readability. For example, we see that on October 22, 2020, the estimated rate of death in Georgia was 0.042, meaning that the average time till death for COVID-19 fatalities was $\frac{1}{0.042} = 23.8$ days. We train TNC and DA using bounds of 20% and 50%, respectively, around the latest parameters trained on the previous day. We utilize the rolling training window of historical data mentioned above, leaving out one week of data as validation. As we observe from the table, the MAPE on the validation set from TNC is lower, and thus the optimized parameters from TNC are accepted, whereas the DA parameters are discarded. After this update, we would continue to use TNC to update the parameter daily until October 30, 2020, when both algorithms are

Table 1. Parameter Update for Georgia on October 23, 2020

Parameter Date	Infection Rate $\tilde{\alpha}$	Rate of Death $\tilde{\tau}$	Resurgence Magnitude \tilde{c}	MAPE
2020-10-22	0.409	0.042	0.924	
2020-10-23 (TNC Bounds)	(0.327, 0.491)	(0.034, 0.050)	(0.739, 1.109)	N/A
2020-10-23 (TNC Result)	0.395	0.036	1.108	1.02%
2020-10-23 (DA Bounds)	(0.205, 0.614)	(0.021, 0.063)	(0.462, 1.386)	N/A
2020-10-23 (DA Result)	0.366	0.021	0.059	1.14%

MAPE refers to the mean average percentage error on the 7-day holdout validation set.

utilized again to select the best-performing parameters. This dual-track approach allows efficient optimization and ensures close data fit. We next discuss three key characteristics of the DELPHI model that allow it to flexibly fit a wide range of time-series of observed data.

2.1. Accounting for Underdetection

Only a subset of the SARS-CoV-2 infections are identified through testing. This is because testing resources were scarce in the beginning of the epidemic; some patients will never develop symptoms that are severe enough to prompt them to seek testing or care, some may attribute any symptoms to another infection, such as the common cold, and some may refrain from getting tested for other reasons (e.g., to avoid losing time from work). Although in reality, the likelihood that a patient will be identified through testing varies across areas and over time, DELPHI treats the probability of detection, p_d , as a global constant *nuisance parameter*. The probability p_d is not identifiable in each area from the available data, namely detected cases and deaths. Fixing it to 20% (which represents average estimates in different countries from the early studies) is, in practice, no different from assuming other reasonable estimates that vary by area. The model has enough degrees of freedom to adjust the values of other area-specific, partially identified parameters (e.g., the area-specific reproduction rate).

Allowing both the detection probability and the reproduction rate to vary by area or over time offers no additional advantage for predicting detected cases and deaths and may occasionally lead to overfitting issues (see, e.g., Lourenço et al. 2020). In sensitivity analyses (Section 3.3), we show that predictions of detected cases and deaths are robust to moderate deviations from the value of $p_d = 20\%$. We also modeled the detection probability as functions of observable data, such as the number of tests that were administered over time in a country, but this did not improve the out-of-sample empirical performance of the model.

DELPHI's goal is not to infer the number of all, detected and undetected, infections. If one wished to point-identify the true detection probability, more and different input data, such as random serology testing, would be required. However, such data were, and still are, very sparse, most often pertaining to specific cities and counties (see Bendavid et al. 2020, Doi et al. 2021, Sood et al. 2020, and Streeck et al. 2020 for examples) and only occasionally to countries (mostly in Europe, (see, e.g., Erikstrup et al. 2020 and Wise 2020).

2.2. Separation of Recovery and Deaths

DELPHI's compartmental structure allows for different transition rates for recovery and death processes

by using the auxiliary compartments U_R , U_D (for undetected patients), DH_R , DH_D (for the hospitalized), and DQ_R , DQ_D (for detected patients), which were described earlier. Figure 1, (b) and (c), explains how this is achieved. In Figure 1(b), the outflow from the I compartment is

$$\frac{dI^-}{dt} = -(v + \tilde{\tau})I,$$

where v is the rate of recovery and τ is the rate of death. This implies a fixed probability of death equal to $\frac{\tilde{\tau}}{v + \tilde{\tau}}$. By contrast, the structure in Figure 1(c) decouples the probability of death $\tilde{\mu}$ from the transition rates $\tilde{\tau}$ and v .

2.3. Modeling Effect of Increasing Government Response

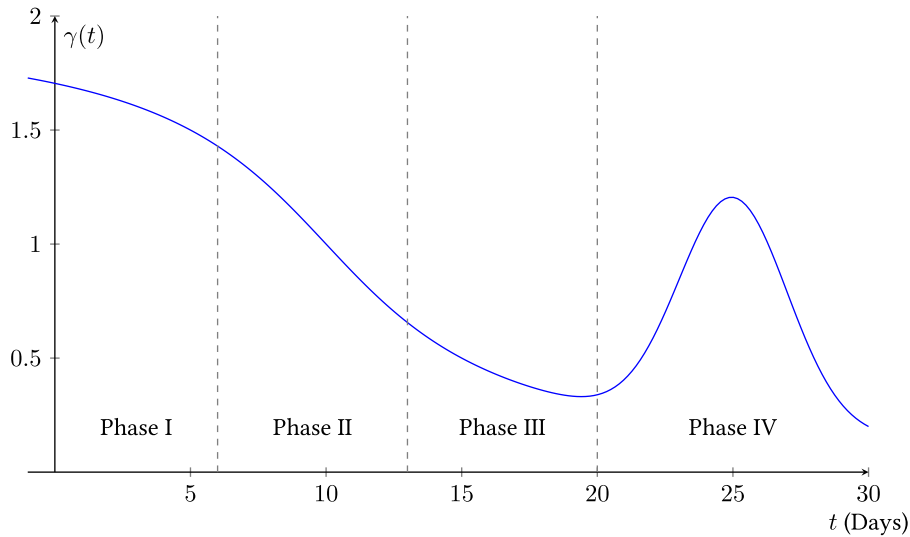
As shown in Figure 2, DELPHI models different phases for the government response during a pandemic by means of the area-specific function $\gamma(t)$ in (1), which includes sigmoid (arctan) and exponential terms. The concave-convex nature of the arctan term accounts for the first three phases. The early, concave part models initial limited changes in behavior in response to early information, when most people continue business-as-usual activities. The transition from the concave to the convex part of the curve quantifies the sharp decline in infection rate as policies go into full force and people's behavior changes sharply. The latter convex part of the curve models a flattening out of the response as the government measures reach saturation, representing the diminishing marginal returns in the decline of infection rate. The exponential term models a potential resurgence in cases, for instance, due to premature relaxation of societal restrictions or some change in behavior.

In (1), parameters \tilde{t}_0 and \tilde{k} control the timing of such measures and the rapidity of their penetration, whereas the $\tilde{c}, \tilde{t}_{\text{jump}}, \tilde{d}$ controls the timing, magnitude, and duration of a resurgence. In Section 4.1, we explain that (1) forms the basis for modeling a wide variety of policies as the composite of elemental interventions, such as including social distancing, school closings, and stay-at-home orders.

2.4. Implementation and Availability

DELPHI was created in early April 2020 and has been continuously updated to reflect new observed data. The codebase is available on GitHub (<https://github.com/COVIDAnalytics/DELPHI>), with the primary model written in Python 3.7 using the SciPy and NumPy libraries. The implementation is also multiprocessing friendly, which allows it to scale easily (using servers/machines

Figure 2. Illustration of the response function $\gamma(t)$ for the particular set of parameters $\tilde{t}_0 = 10, \tilde{k} = 5, \tilde{c} = 1, \tilde{t}_{\text{jump}} = 25,$ and $\tilde{d} = 2$ (i.e., $\gamma(t) = 1 + \frac{2}{\pi} \arctan\left(-\frac{t-10}{5}\right) + \exp\left(-\frac{(t-25)^2}{8}\right)$).



with multiple CPUs/threads) to the high number of areas the model is fitted on every day.

3. Results and Performance Analysis

In this section, we present the results of the DELPHI predictive model and its performances in terms of mean absolute percentage error (MAPE) and root mean squared error (RMSE) across time and regions and benchmark it against the state-of-the-art COVID-19 models used by the CDC. We also analyze the sensitivity of DELPHI to perturbations in its parameters.

3.1. Forecasting Results

Table 3 reports the median MAPE and RMSE for the observed cumulative numbers of cases and deaths in each area of the world for two periods. The first uses data through April 27, 2020, and evaluates models up until May 12, 2020. The second uses data up to September 21, 2020, and evaluates models through October 6, 2020. During the second period, there was a resurgence of the epidemic, the management of the disease was better understood, and new treatment protocols were in place. For both periods, DELPHI seems to predict the epidemic progression relatively well in most areas, with < 10% MAPE on reported cases and < 15% MAPE

on reported deaths. The worldwide median MAPE was 5.8% for detected cases and 10.6% for deaths. The areas with the highest MAPE were typically those with the fewest deaths, as shown in the selected examples in Table 2. Analogously, the median RMSE for deaths in both periods was <100 across all regions, which is remarkable given that the RMSE is not scaled by the observed number of deaths and the high number of deaths per region (e.g., by the second period, a majority of areas in North America and Europe were reporting more than 5,000 cumulative deaths). The median RMSE on cases is also remarkable, at just above 2,000 cases in the second period, given the daily median number of recorded cases across all areas (~ 60,000).

The worldwide median MAPE for deaths in the second period is smaller than in the first (4.8% vs. 10.8%, respectively). This may be partially explained by the better fitting of the probability of death $\tilde{\mu}(t)$ in the second period, when the cumulative number of deaths was higher and also the higher variability of MAPE when the denominator (observed deaths) is small (analogously to the observation in Table 2).

We now further illustrate the performance of DELPHI with two major countries with very different curves. Figure 3, (a) and (b), shows our projections of the number of cases in Russia and the United Kingdom

Table 2. Breakdown of Cumulative Number of Deaths vs. Corresponding Prediction MAPE for Large Errors on the Prediction Period of April 28 to May 12

Country/Province	Bahrain	Djibouti	Guinea	Kazakhstan	Sri Lanka	Oman	Qatar	Venezuela
Cumulative deaths as of May 11, 2020	8	3	11	32	9	17	14	10
MAPE on deaths	89.6%	193.1%	53.3%	62.6%	54.5%	44.0%	106.9%	48.1%

Table 3. Median Country-Level Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) of the Predicted Number of Cases and Deaths in Each Region

Region	No. of Areas	Median MAPE Cases (10th, 90th percentile)	Median MAPE Deaths (10th, 90th percentile)	Median RMSE Cases (10th, 90th percentile)	Median RMSE Deaths (10th, 90th percentile)
<i>April 28th</i>					
Africa	19	14.7% (3.1, 32.0)	23.4% (11.8, 60.3)	138.7 (27.0, 1019.6)	4.8 (1.1, 27.5)
Asia	32	4.8% (2.1, 18.4)	14.4% (2.9, 65.2)	677.3 (51.5, 9778.1)	13.0 (1.0, 151.6)
Europe	42	3.4% (0.8, 12.9)	9.0% (2.3, 24.3)	238.1 (15.4, 3276.4)	14.8 (1.6, 236.2)
North America	10	7.9% (3.9, 28.3)	12.6% (2.8, 23.6)	594.4 (36.1, 1947.5)	16.3 (4.0, 132.3)
Oceania	2	3.2% (2.4, 4.1)	12.0% (11.0, 13.0)	68.6 (49.4, 87.8)	2.0 (1.8, 2.3)
South America	11	14.9% (7.6, 26.7)	6.1% (3.3, 30.1)	683.6 (31.4, 10815.0)	6.8 (0.6, 426.1)
United States	51	8.5% (1.9, 16.7)	7.8% (3.3, 25.1)	1231.6 (73.8, 4861.8)	33.5 (1.7, 210.7)
World	167	5.8% (1.5, 22.6)	10.6% (2.9, 36.6)	412.6 (26.1, 4788.7)	12.0 (1.3, 193.1)
<i>September 22nd</i>					
Africa	53	5.2% (0.6, 30.4)	4.2% (0.0, 42.6)	364.5 (27.5, 2913.2)	8.8 (0.0, 121.7)
Asia	38	6.5% (1.7, 38.4)	8.3% (1.4, 26.8)	6311.3 (120.9, 53046.7)	47.4 (0.7, 719.5)
Europe	44	13.4% (3.8, 38.5)	7.7% (1.1, 22.5)	4452.9 (326.6, 22405.9)	43.6 (1.7, 776.5)
North America	14	7.3% (1.2, 17.1)	4.8% (1.0, 22.0)	2180.4 (187.5, 8656.7)	48.9 (2.0, 269.0)
Oceania	3	2.7% (1.6, 4.2)	1.1% (0.8, 7.1)	51.0 (16.0, 127.5)	0.8 (0.6, 3.8)
South America	13	9.4% (1.2, 15.5)	5.2% (2.1, 22.7)	5683.8 (473.2, 131184.4)	95.6 (3.5, 4438.1)
US	51	5.2% (1.3, 20.6)	3.0% (0.5, 15.0)	5323.5 (661.6, 17571.2)	66.9 (3.1, 275.0)
World	216	6.5% (1.2, 28.2)	4.8% (0.6, 27.6)	2170.0 (69.3, 18549.4)	31.2 (1.1, 505.3)

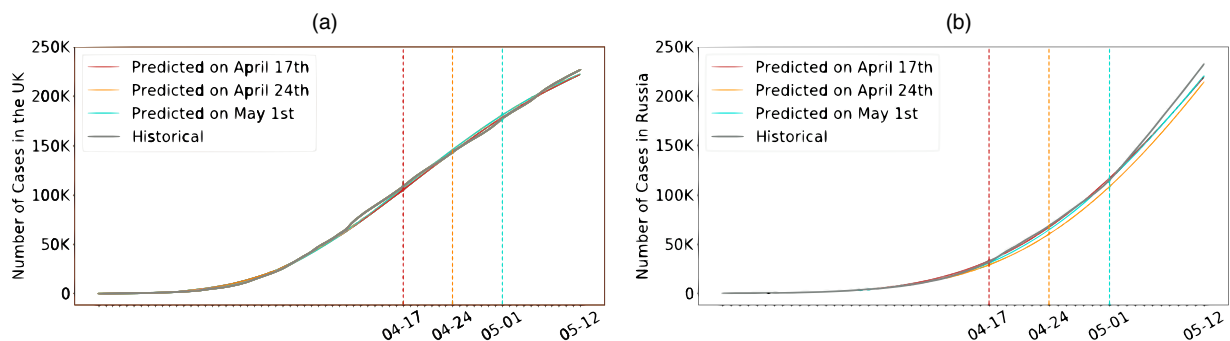
Projections in the first (resp. second) half of the table are made using data up to 04/27 (resp. 09/21) for the period from 04/28 to 05/12 (resp. 09/22 to 10/06).

made on three different dates and compares them against historical observations. The results for Russia and the United Kingdom are consistent with the overall performances across all countries of all regions, as described extensively in the rest of the section. Concretely, the graphs suggest that DELPHI achieves strong predictive performance, because the model has been consistently predicting, with high accuracy, the overall spread of the disease for several weeks across regions with different epidemiological characteristics. Notably, DELPHI was able to anticipate, as early as April 17, the dynamics of the pandemic in the United Kingdom (resp. Russia) up to May 12. At a time when 100,000 – 110,000 (resp. 30,000 – 35,000) cases were reported, the model was predicting 220,000 – 230,000 (resp. 225,000 – 235,000) cases by May 12, a prediction

that was realized a month later. In the case of Russia, DELPHI was able to predict that the country was going to become a global hotspot as well as to accurately estimate the magnitude of the first wave of the outbreak, even at an early stage of the pandemic (less than 0.025% of the population infected, versus more than 0.16% a month later, which has put the country at the 4th rank worldwide in terms of cumulative number of cases).

3.2. Comparison With Other Models

DELPHI compares favorably with other top-performing models submitted to the CDC ensemble forecast in predicting the number of deaths in the United States four weeks ahead (the longest time point in the CDC ensemble predictions). As comparator models we selected

Figure 3. Cumulative Number of Cases in the United Kingdom (a) and Russia (b) According to Our Projections Made at Different Points in Time Against Actual Observations

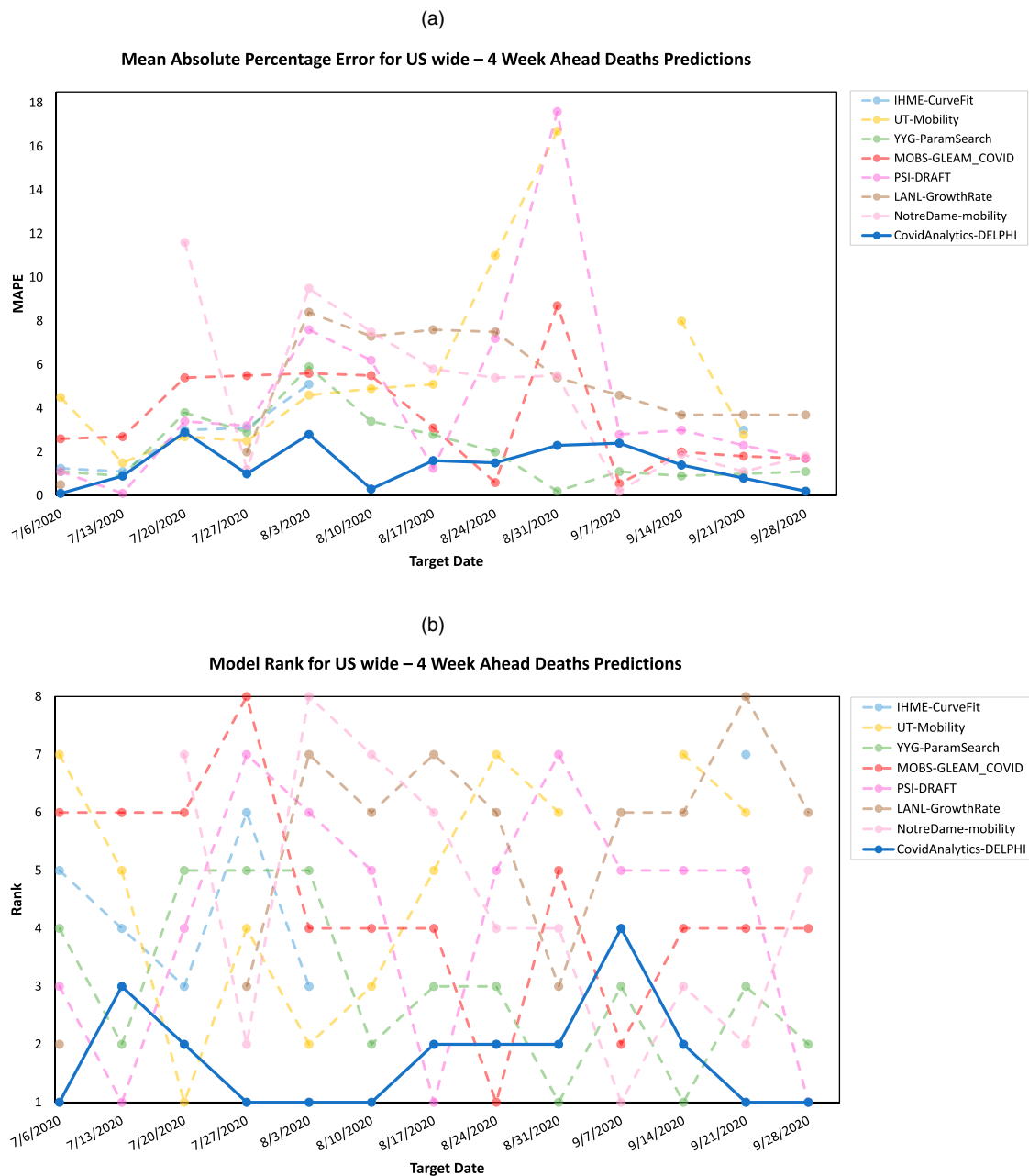
Note: There predicted curves largely overlap with the actual curve: (a) United Kingdom; (b) Russia.

those that submitted results to the CDC regularly during the epidemic and were thus consistently included in the CDC ensemble forecasts. These were the models by the University of Texas, Austin (UT-Mobility, Woody et al. 2020), the Institute for Health Metrics and Evaluation (IHME COVID-19 Health Service Utilization Forecasting Team and Murray 2020), Youyang Gu (Gu 2020), the Northeastern University’s Laboratory for the Modeling of Biological and Socio-technical Systems

(MOBS-GLEAM_COVID, Chinazzi et al. 2020), Predictive Science, Inc. (PSI-DRAFT 2020), the Los Alamos National Laboratory (LANL COVID-19 Team 2020), and the Notre Dame University (NotreDame-mobility, Perkins and Espana 2020).

In Figure 4(a), we compare the out-of-sample MAPE of these models for the number of cumulative deaths four weeks in the future above using the actual weekly predictions submitted to the CDC between July and

Figure 4. Comparison of 4-Week MAPE on Deaths Prediction in the United States Between DELPHI and Other Models Used by the CDC



Note: (a) MAPE of DELPHI for United States wide 4-week ahead deaths prediction from July to September; (b) rank of DELPHI for United States wide 4-week ahead deaths prediction from July to September.

September 2020. This particular period was selected because it encompassed the period of second resurgence in the United States and its decay, making prediction even more difficult. October 2020 was excluded because the CDC ensemble forecast changed its submission and reporting guidelines, prompting submission lapses in many models. We observe that DELPHI consistently achieves low MAPE and that its predictions are stable with a MAPE, never exceeding 3.5% throughout the period. Figure 4(b) further illustrates the performance of DELPHI in comparison with other models by graphing the weekly ranking (with respect to MAPE). We observe that DELPHI consistently outperforms all other models, holds the first rank for six out of 13 weeks, and never drops below rank 4 among the eight models evaluated.

3.3. Sensitivity Analysis

We examined the impact on prediction of varying each of the six fixed parameter of DELPHI in univariate sensitivity analyses. For every fixed parameter among β , r_d , σ , κ , p_d , and p_h , we randomly perturbed the parameter by a zero-centered normal noise term ε with standard deviation of 20% of the nominal parameter's absolute value, that is, $\varepsilon \sim \mathcal{N}(0, (0.2 \cdot |\text{param}|)^2)$. Then we fit the DELPHI model using data up to a certain prediction date using the perturbed fixed parameter and compared its 30-day out-of-sample MAPE with that of the baseline value.

We conducted these sensitivity analyses for all states in the United States and in six countries around the world with large outbreaks (Italy, Spain, Brazil, South Africa, Japan, and Russia) for three prediction dates.

Figure 5, (a) and (b), records the quantile (box and whisker) plots of the absolute difference between the MAPE of the actual model and the perturbed model for six parameters ($\beta, r_d, \sigma, \kappa, p_d, p_h$), across the 56 areas (50 U.S. states and six countries) for the three prediction dates. We observe that for all six parameters, across both cases and deaths, the effect of the perturbation on the one-month MAPE is relatively small, with interquartile range mostly falling between $\pm 5\%$ for a perturbation with a standard deviation of 20% of the parameter value. This demonstrates that the results from the DELPHI model are robust to a moderately large perturbation to the underlying parameters.

4. Applications

DELPHI's predictions of the epidemic's trajectory can inform decisions of policymakers and research design, as shown in two selected applications. In the first application, we consider different NPIs to limit social interactions and mixing and extend the DELPHI model to evaluate their impact on the trajectory of the epidemic. The second application demonstrates a scenario analysis toolkit that can inform the

planning of operations, staffing, inventories, or even the development of research designs. This toolkit was used by Janssen Pharmaceuticals to select candidate sites for their Phase III trial of the single-dose COVID-19 vaccine Ad26.Cov-2.S.

4.1. Application 1: Evaluating Different Government Intervention Scenarios

In this section, we extend DELPHI to evaluate the impact of government interventions. That allows us to quantify the efficiency of NPIs and predict "what if" scenarios under different policies, which enable policy-makers to assess their COVID-19 response and decide on their future interventions accordingly. We begin by focusing on the effect of different interventions and then analyze different "what if" scenarios.

4.1.1. Effect of Government Interventions. We can use DELPHI to examine the association between five escalating policy categories (1. no measure; 2. restrict travel and work; 3. restrict mass gatherings, travel and work; 4. restrict mass gatherings, schools, travel and work; and 5. stay at home) and the daily infection rate across areas during the first training window up to May 19, 2020. For each area, we assign each day in the first training window to one of the five policy categories using data from the Oxford Coronavirus Government Response Tracker (Hale et al. 2021) for countries other than the United States and the Institute for Health Metrics and Evaluation (Murray et al. 2020) for U.S. states (detailed correspondence in the e-companion).

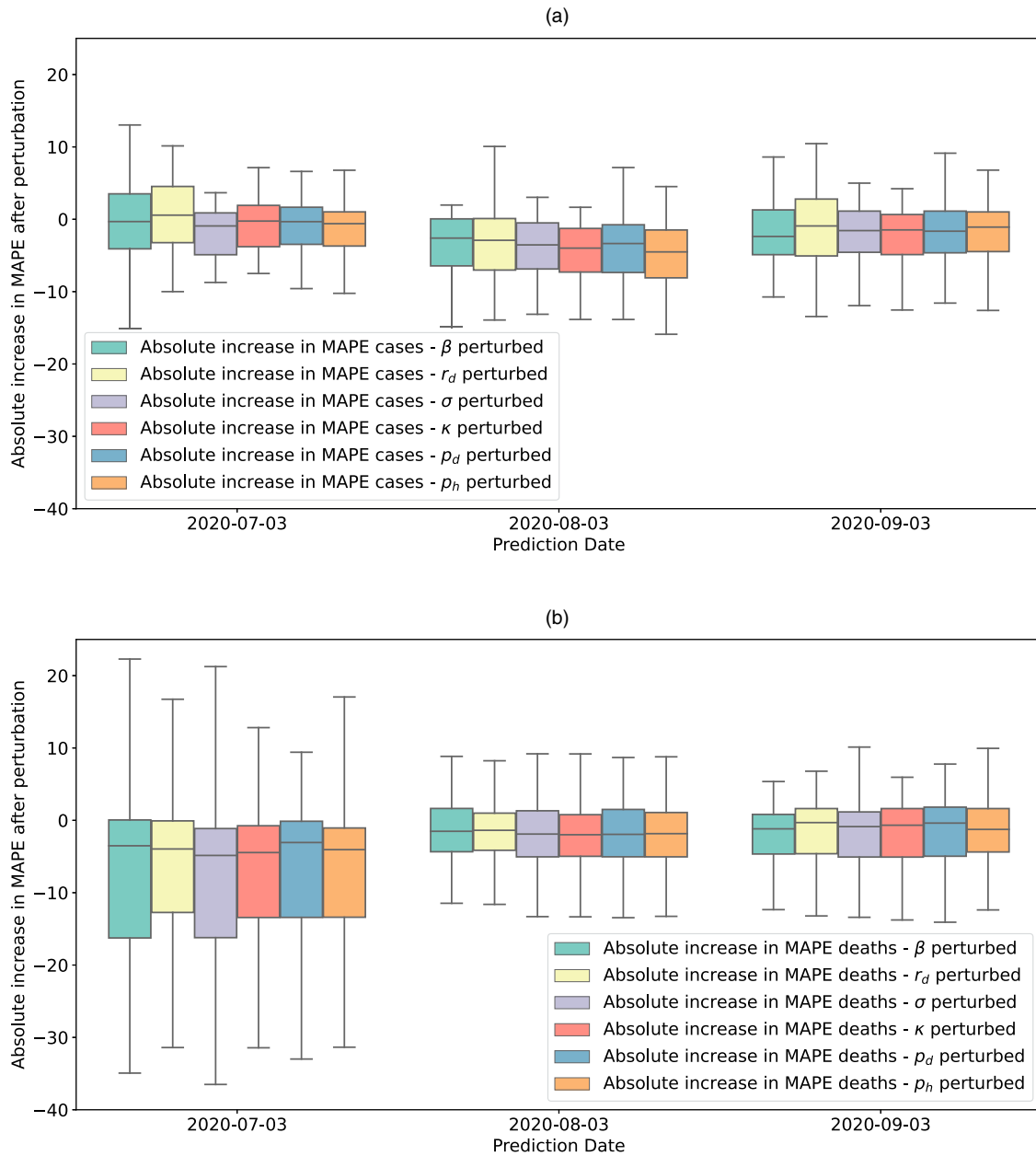
We assume that each of the five policy categories has a global fixed effect across all areas. We use a two-step approach. First, we estimate the $\gamma(t)$ function for each area and training window, which captures the net effect of everything that affects infection rates, including government interventions, changes in behavior, workplace policies, etc. In the second step, we treat the $\gamma(t)$ functions as known and use the presence or absence of the aforementioned five policy categories to explain temporal changes in area-specific infection rates using the algorithm below.

For each policy category $i = 1, \dots, 5$ and each area j , we extract the average value of $\gamma(t)$, $\bar{\gamma}_{ij}$ across all times for which policy i was in effect. Then we calculate the residual fraction of infection rate under policy i , p_i , compared with the baseline policy of no measure as

$$p_i := \frac{1}{J} \sum_{j=1}^J p_{ij} = \frac{1}{J} \sum_{j=1}^J \frac{\bar{\gamma}_{ij}}{\bar{\gamma}_{1j}}$$

where J is the total number of areas we include and p_{ij} is the residual fraction of infection rate under policy i for a specific area j . We normalize the residual fraction of infection rates for different areas because different areas have different background infection rates $\tilde{\alpha}$.

Figure 5. Sensitivity Analysis of Various Fixed Parameters, Comparing Perturbed MAPE on Cases and Deaths to their Nominal Counterparts Without Perturbations



Note: (a) Sensitivity of predictions on cases based on perturbation of key fixed parameters; (b) sensitivity of predictions on deaths based on perturbation of key fixed parameters.

The estimated “effects” of the five policy categories on infection rates are associations rather than causal estimates in that they may be confounded by other unmeasured or unobservable factors. For example, during the first wave, most governments had not yet implemented a mask mandate, and thus we did not include masks in the policy categories. However, some did, and the effects of masking are aliased with the effects of the five policies. On the other hand, the average effect of unaccounted for policies that are active throughout the epidemic, such as contact tracing, is

absorbed by the area-specific parameter $\tilde{\alpha}_j$, which cancels out when estimating each $p_{ij} = \frac{\tilde{y}_{ij}}{\tilde{y}_{1j}}$. This suggests a small impact of time-invariant unmeasured policies on the estimates p_i .

Table 4 shows the number of area days that each policy was implemented around the world and its average effect over all areas and the standard deviation of the area-specific estimates p_{ij} for each policy i . During the first wave, each policy category was in effect for hundreds to thousands of area days worldwide, with

Table 4. Implementation Length and Effect of Each Policy Category as Implemented Across the World

Restrictions	Area-Days	Residual Infection Rate
None	2142	100%
Travel and Work	2049	88.9 ± 4.5%
Mass Gathering, Travel, and Work	340	59.0 ± 5.2%
Mass Gathering, School, Travel, and Work	1460	41.7 ± 4.3%
Stay-at-Home Order	6585	25.6 ± 3.7%

the stringent stay-at-home policy category being implemented most extensively. Table 4 lists policies in increasing estimated effectiveness. Compared with less restrictive policies, more-stringent policies tend to be implemented later and associated with larger decreases in the residual infection rate. This is expected, because during the first wave the $\gamma(t)$ function is monotonically decreasing ($\bar{c} = 0$), and later-implemented policies will have larger estimated effects. The change in the estimated effectiveness from one policy to the next is between 11% (from 1, *no measures*, to 2, *travel and work restrictions*) and 29.9% (from 2 to 3, *mass gathering, school, travel, and work restrictions*). The most stringent policy category (5, *stay at home*) is associated with a reduction of the infection rates to $25.6 \pm 3.7\%$ of the unmitigated value.

The basic reproduction rate, R_0 , for the SARS-CoV-2 variant responsible for the first wave of COVID-19 was estimated to be between 2.5 and 3.0 (Liu et al. 2020a, Zhang et al. 2020). R_0 measures on average how many new infections one infected patient will generate over the course of their disease in a fully susceptible population (i.e., in the beginning of the epidemic and, approximately, during the first wave). The basic reproduction rate R_0 is proportional to $\gamma(0)$, and the effective reproduction rate R_t , its counterpart for $t > 0$, is proportional to $\gamma(t)$. To control the epidemic R_t should become smaller than 1, or $\frac{\gamma(0)}{\gamma(t)} \lesssim \frac{1}{2.5}$ to $\frac{1}{3.0}$, if at $t = 0$ the epidemic was unmitigated. This suggests that, to the extent that $\frac{\gamma(0)}{\gamma(t)} \approx \frac{\bar{\gamma}_t}{\bar{\gamma}_1} = p_i$, on average, only policy $i = 5$, stay-at-home-order, appears to be strong enough to fully mitigate the epidemic, albeit at a steep economic and social cost. As we will discuss below, the timing when policies go into effect is critical for minimizing the total number of cases and deaths.

4.1.2. Modeling Alternative Initial Responses. To model what would happen if mitigation were to start m days earlier, we translate the time axis in (1) by m days to the left:

$$\gamma'(t) = \frac{2}{\pi} \arctan\left(-\frac{t - (\tilde{t}_0 - m)}{\tilde{k}}\right) + 1.$$

To illustrate, Figure 6 shows the percentage of cases and deaths avoided around the world by May 17 if the government interventions were to be initiated one week earlier for the 50 countries with the highest reduction, which ranges between approximately 30% and 80%. Western European countries such as Switzerland, Spain, and Italy, which had some of the first and steepest outbreaks outside Asia, would have benefited the most. Cumulatively across the world, DELPHI predicts that more than 280,000 deaths or 68% of total deaths could have been avoided by May 17 with just a week's earlier start of mitigation efforts.

Another insightful scenario to consider is what would have happened if the epidemic were left unmitigated. This can be modeled as

$$\gamma'(t) = \gamma(0) = \frac{2}{\pi} \arctan\left(\frac{\tilde{t}_0}{\tilde{k}}\right) + 1,$$

and would result in more than 14.8 million deaths by May 17, 2020.

4.2. Application 2: Analysis of What If Scenarios for Long-Term Planning

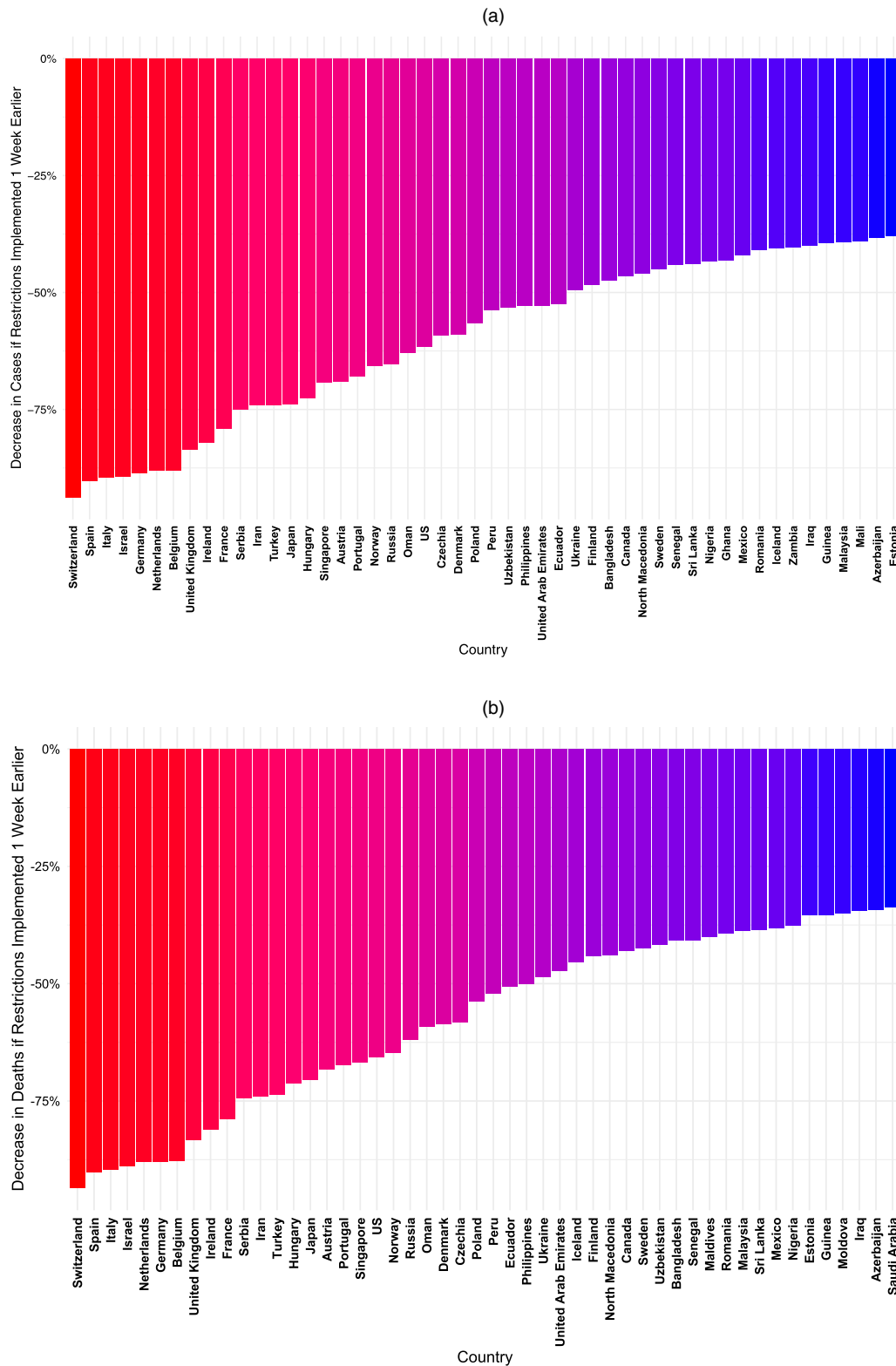
DELPHI was utilized by Janssen Pharmaceuticals in late May 2020 to examine the impact of “what if” scenarios of relaxing measures in different countries to inform the design of the multicenter Phase III trial of their vaccine candidate Ad26.Cov-2.S. DELPHI's predictions helped identify the best candidate sites (countries with high anticipated incidence and prevalence) to maximize the trial's statistical power.

Specifically, suppose that we are considering shifting from policy i to $j < i$ at time t_c in some area that has not yet experienced a resurgence ($\bar{c} = 0$). Then, for all times $t \geq t_c$, we modify (1) to

$$\gamma'(t) = \frac{2}{\pi} \arctan\left(-\frac{t - \tilde{t}_0}{\tilde{k}}\right) + 1 + \underbrace{(p_j - p_i) \cdot \min\left[\frac{2 - \gamma'(t_c)}{1 - p_i}, \frac{\gamma'(t_c)}{p_i}\right]}_{\text{Differential in policy effect between policy } i \text{ and } j}, \quad \forall t \geq t_c.$$

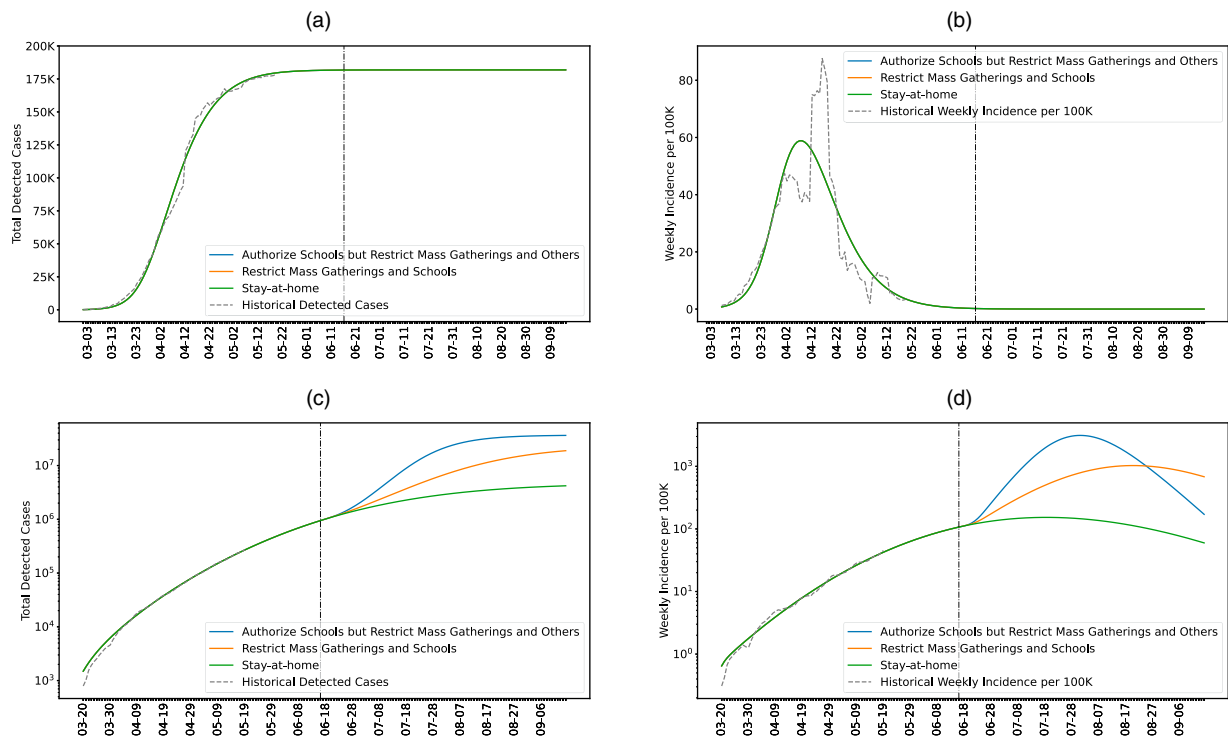
The last term is a correction proportional to the difference $p_i - p_j > 0$ in the fractional reductions in the

Figure 6. Scenario analysis if restrictions implemented one week earlier



Note: (a) Percentage of cases avoided around the world if policy enacted one week early; (b) percentage of deaths avoided around the world if policy enacted one week early.

Figure 7. Forecasts of Total Detected Cases and Weekly Incidence Per 100K for France and Brazil Under Various Policies



Notes: In (a) and (b), the green line completely overlaps with other lines in the asymptotic regime. (a) France, total Detected Cases; (b) France, weekly incidence per 100K; (c) Brazil, total detected cases (log scale); (d) Brazil, weekly incidence per 100K (log scale).

infection rate with policy categories i and j . The multiplicative factor $\min\left[\frac{2-\gamma(t_c)}{1-p_i}, \frac{\gamma(t_c)}{p_i}\right]$ scales the fractional difference so that the resulting $\gamma'(t_c)$ is constrained within the initial range $[0, 2]$. Replacing $\gamma(t)$ with $\gamma'(t)$ forecasts the epidemic under the updated policy.

The impact of a reopening strategy varies greatly across areas, as shown in a comparison of Brazil versus France in Figure 7. In Brazil, relaxing measures from a stay-at-home order (policy category 5) to restricting mass gathering, travel, and work (category 3) on June 16 would result in a second wave of infections, with up to 6.8 million additional cases one month later (Figure 7(c)), because the epidemic was not yet adequately mitigated and the incident cases were still on a steep rise (Figure 7(d)). By contrast, in France, the epidemic had already peaked and was adequately mitigated, incident detected cases were declining (Figure 7(b)), and the relaxation of policies would have a much smaller effect (Figure 7(a)). Results for other countries are shown in the e-companion.

Figure 8 summarizes analogous scenario analyses across the globe. It shows outcomes one month after a hypothetical relaxing of policy category 5 (stay at home) to 3 (mass gathering, travel, and work restrictions) on June 16, 2020. We observe three clusters of countries:

- Countries with relatively few total cases, with good mitigation of the epidemic (with relatively few

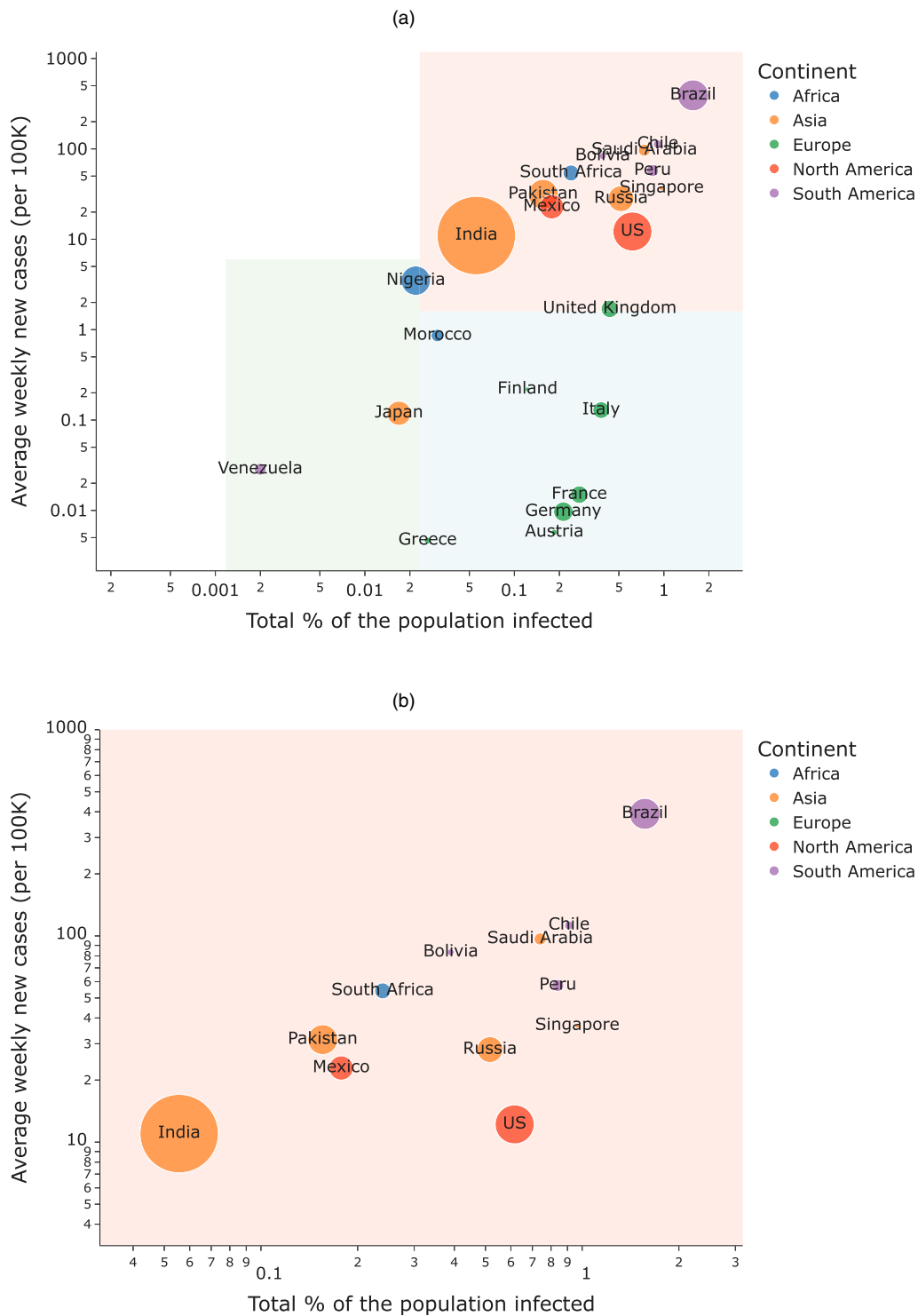
incident cases), such as Greece, Japan, Morocco, and Venezuela,

- Countries with comparatively higher total cases, also with an adequate mitigation of the epidemic, mainly in Western and Northern Europe (e.g., the United Kingdom, Italy, France, and Finland), and

- Countries with a large and relatively fast-growing number of cumulative cases, including the United States, India, and Brazil, in which the epidemic is not adequately mitigated. A close-up of these countries is presented in Figure 8(b). For example, DELPHI predicted that up to 8% of Brazil’s population would be confirmed with COVID-19 one month after a relaxation to policy category 3.

Janssen Pharmaceuticals applied this analysis around the world in May–June 2020 to identify candidate sites for their Ad26.Cov2.S vaccine trial, which was planned for September 2020. For all candidate countries, they performed analyses for all policy relaxations and prioritized for further consideration and feasibility analyses those with predicted weekly incidence of confirmed cases >25 per 100,000 people by the anticipated trial start date. For example, from Figure 7, Brazil would be a candidate country, but France would not be. Such analyses informed their final selection of eight countries, namely, Argentina, Brazil, Chile, Columbia, Mexico, Peru, South Africa, and the United States.

Figure 8. World Predictions for Early July Under Mass Gathering, Travel, and Work Restrictions



Note: (a) Weekly incidence of cases (per 100K) in the first half of July against fraction of population infected for multiple countries; (b) predictions for total cumulative cases (normalized by the population) vs. new cases (per 100K) for countries which are predicted to be highly impacted and still worsening at an alarming rate by July 15th.

Notably, prior to using DELPHI, Janssen did not consider Brazil and South Africa. In retrospect, adding these two high-incidence countries provided valuable

information on the effectiveness of the vaccine against emerging SARS-CoV-2 variants (gamma in Brazil, beta in South Africa).

5. Limitations

We briefly discuss several limitations of our approach.

First, DELPHI is a deterministic model, with several input parameters fixed to literature-derived point estimates and other parameters fit to data. As presented here, it is not used for uncertainty propagation and quantification tasks (Council et al. 2012), which are important for contextualizing forecasts and predictions of “what if” scenarios for policymaking. However, insights from deterministic models are still useful in understanding the dynamics of the disease and for practical uses, as demonstrated in the applications. Changing the input parameters from fixed values (point mass distributions) to parametric or empirical probability models is straightforward but would be computationally intensive, complicating logistics.

Second, all input parameters that were not fit to data were fixed to global values obtained from the early literature. These include the probability that a case is detected (p_d) and parameters that describe the biology of the disease, such as the mean duration of the incubation period (β^{-1}), the mean time to detection (r_d^{-1}), and the probability that an infection will result in hospitalization (p_h). Some of these parameters, such as the mean time to detection, are fairly consistent in the literature (see, e.g., Grein et al. 2020, Hu et al. 2020, Kluytmans et al. 2020, Lauer et al. 2020, Liu et al. 2020b). For others, information is sparse. For example, the true (unobserved, latent) detection probability varies by country and over time. However, it is not identifiable in each modeled area on the basis of the available data, namely, detected cases and deaths. Fixing it to 20% (which represents an average of early estimates in different countries from the early studies) is, in practice, no different from assuming other reasonable estimates that vary by area. The model has enough degrees of freedom to adjust the values of other area-specific, partially identified parameters (e.g., the area-specific reproduction rate). With DELPHI, the goal is to predict future detected cases and deaths. If one wished to also calibrate the model so that it could infer (fit) the actual detection probability, more and different input data would be required, for example, random serology testing data. However, such data were (and still are) only very sparsely available among the 200+ areas in which we make predictions.

Third, DELPHI does not explicitly account for population stratification by sex, age, or occupation. There can be substantial variation in transmissions within and across population strata, especially age groups, which in turn can give rise to complicated dynamics (Larson 2007, Britton et al. 2020, Gomes et al. 2022). However, marginal modeling of the population allowing some quantities to vary over time (e.g., here $\gamma(t)$, $\mu(t)$) suffices to approximate any dynamics induced by population heterogeneity. This is demonstrated empirically by the

fact that DELPHI (and other marginal models) capture the realized dynamics across U.S. states (Dean et al. 2020) and in many countries.

Fourth, DELPHI uses a marginal probability of death ($\mu(t)$) (infection fatality rate) and a marginal mean time to death $\tilde{\tau}^{-1}$ for all compartments, irrespective of hospitalization or detection status. However, it has enough degrees of freedom to satisfactorily predict the total detected cases and deaths in diverse settings and under varying mitigation policies.

Fifth, DELPHI assumes that patients participate in the infection process only for an average duration of $r_d^{-1} \approx 2.9$ days, after the disease has incubated and before patients are self-quarantined (for undetected cases) or quarantined or hospitalized (for detected cases). However, because $R_0 = \tilde{\alpha} \cdot \gamma(0) \cdot r_d^{-1}$, only the ratio $\tilde{\alpha}/r_d$ is identified, and the fitting algorithm simply finds a suitable area-specific infection rate $\tilde{\alpha}$. Furthermore, DELPHI assumes that all patients will eventually develop enough symptoms to prompt their self-isolation or to seek care and get test-detected. Although currently it is understood that not all patients develop symptoms, that asymptomatic transmission is possible but perhaps less common than during symptoms (Byambasuren et al. 2020, Ing et al. 2020, Sayampanathan et al. 2021), and that some patients with symptoms will not self-isolate or seek detection or care (e.g., to not lose working days and income), this was not as clear early on. However, as discussed above, DELPHI has enough degrees of freedom to yield empirically good predictions.

Finally, there are shortcomings in our estimation of the effects of policy interventions in the first application. As discussed in the first application, the estimated “effects” are associations rather than causally interpretable quantities, and they can be confounded by other, unmeasured or unobservable, factors. To facilitate fitting, we assumed that the relative reduction in the infection rate p_i associated with the i -th policy category is homogeneous across areas. These simplifications were motivated by lack of data and pertinent information. We believe that the key qualitative insights on the differential impact of relaxation of policies across areas and the impact of earlier versus later implementation of measures are fundamental and robust to these simplifications.

6. Conclusions

DELPHI is a detailed epidemiological model that accurately predicted the spread of COVID-19 in many countries and aided planning for many organizations worldwide, including governmental entities, hospitals, and pharmaceutical companies. By modeling the impact of government interventions, DELPHI provided key insights on the effects of differential timing

in the implementation and relaxations of government policies on total detected cases and deaths.

References

- Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, Taylor J, et al. (2020) Presymptomatic Sars-CoV-2 infections and transmission in a skilled nursing facility. *N. Engl. J. Med.* 382:2081–2090.
- Bendavid E, Mulaney B, Sood N, Shah S, Ling E, Bromley-Dulfano R, Lai C, Weissberg Z, Saavedra-Walker R, Tedrow J, Bogan A, Kupiec T, Eichner D, Gupta R, Ioannidis JP, Bhattacharya J. (2020) COVID-19 antibody seroprevalence in Santa Clara County, California. *Int. J. Epidemiol.* 50(2):410–419.
- Bertsimas D, Bandi H, Boussioux L, Cory-Wright R, Delarue A, Digalakis V, Gilmour S, et al (2020) An aggregated data set of clinical outcomes for COVID-19 patients. http://www.covidanalytics.io/dataset_documentation.
- Britton T, Ball F, Trapman P (2020) A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* 369(6505):846–849.
- Byambasun O, Cardona M, Bell K, Clark J, McLaws ML, Glasziou P (2020) Estimating the extent of true asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. *JAMMI* 5(4):223–234.
- Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, Piontti AP, Mu K, Rossi L, Sun K, Viboud C, Xiong X, Halloran ME, Longini IM Jr, Vespignani A (2020) The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* 368(6489):395–400.
- National Research Council (2012) *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. (National Academies Press; Washington, D.C.).
- Dean NE, Piontti AP, Madewell ZJ, Cummings DA, Hinchings MD, Joshi K, Kahn R, Vespignani A, Halloran ME, Longini IM Jr (2020) Ensemble forecast modeling for the design of COVID-19 vaccine efficacy trials. *Vaccine* 38(46):7213–7216.
- Doi A, Iwata K, Kuroda H, Hasuike T, Nasu S, Kanda A, Nagao T, Nishioka H, Tomii K, Morimoto T, Kihara Y (2021) Estimation of seroprevalence of novel coronavirus disease (COVID-19) using preserved serum at an outpatient setting in Kobe, Japan: A cross-sectional study. *Clin. Epidemiol. Global Health* 11:100747.
- Erikstrup C, Hother CE, Pedersen OBV, Mølbak K, Skov RL, Holm DK, Saekmose S, Nilsson AC, Brooks PT, Boldsen JK, Mikkelsen C, Gybel-Brask M, Sørensen E, Dinh KM, Mikkelsen S, Møller BK S, Haunstrup T, Harritshøj L, Jensen BA, Hjalgrim H, Lillevang ST, Ullum H (2020) Estimation of SARS-CoV-2 infection fatality rate by real-time antibody screening of blood donors. *Clin. Infect. Dis.* 72(2):249–253.
- Gomes MGM, Corder RM, King JG, Langwig KE, Souto-Maior C, Carneiro J, Gonçalves G, Penha-Gonçalves C, Ferreira MU, Aguas R (2022) Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold. *J. Theor. Biol.* 540:111063.
- Grein J, Ohmagari N, Shin D, Diaz G, Asperges E, Castagna A, Feldt T, Green G, Green ML, Lescure FX, Nicastri E, Oda R, et al. (2020) Compassionate use of remdesivir for patients with severe COVID-19. *N. Engl. J. Med.* 382:2327–2336.
- Gu Y (2020) COVID-19 projections using machine learning. Retrieved October 10, <https://covid19-projections.com/>.
- Hale T, Angrist N, Goldszmidt R, Kira B, Petherick A, Phillips T, Wester S, Cameron-Blake E, Hallas L, Majumdar S, Tatlow H (2021) A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* 5: 529–538.
- Hu Z, Song C, Xu C, Jin G, Chen Y, Xu X, Ma H, et al (2020) Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. *Sci. China Life Sci.* 63:706–711.
- IHME COVID-19 Health Service Utilization Forecasting Team, Murray CJL (2020) Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. Preprint, submitted March 30, <https://doi.org/10.1101/2020.03.27.20043752>.
- Ing AJ, Cocks C, Green JP (2020) COVID-19: in the footsteps of Ernest Shackleton. *Thorax* 75(8):693–694.
- Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond., A Contain. Pap. Math. Phys. Character.* 115(772):700–721.
- Kluytmans M, Buiting A, Pas S, Bentvelsen R, van den Bijllaardt W, van Oudheusden A, van Rijen M, Verweij J, Koopmans M, Kluytmans J (2020) SARS-CoV-2 infection in 86 healthcare workers in two Dutch hospitals in March 2020. Preprint, submitted March 31, <https://doi.org/10.1101/2020.03.23.20041913>.
- Krantz SG, Rao ASR (2020) Level of under-reporting including under-diagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. *Infect. Control Hosp. Epidemiol.* 41(7):857–859.
- LANL COVID-19 Team (2020) LANL COVID-19 cases and deaths forecasts. Retrieved October 10, <https://covid-19.bsvgateway.org/>.
- Larson RC (2007) Simple models of influenza progression within a heterogeneous population. *Oper. Res.* 55(3):399–412.
- Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich NG, Lessler J (2020) The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Intern. Med.* 172(9):577–582.
- Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J (2020a) The reproductive number of COVID-19 is higher compared with SARS coronavirus. *J. Travel Med.* 27(2):taaa021.
- Liu Y, Sun W, Li J, Chen L, Wang Y, Zhang L, Yu L (2020b) Clinical characteristics and progression of 2019 novel coronavirus-infected patients concurrent acute respiratory distress syndrome. Preprint, submitted February 27, <https://www.medrxiv.org/content/10.1101/2020.02.17.20024166v3>.
- Lourenço J, Paton R, Ghafari M, Kraemer M, Thompson C, Simmonds P, Klennerman P, Gupta S (2020) Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. Preprint, submitted March 26, <https://doi.org/10.1101/2020.03.24.20042291>.
- Mehrotra P, Ivan J (2020) Prophet logistic forecasting.
- Murray C, et al (2020) Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the usa and european economic area countries. Preprint, submitted April 26, <https://doi.org/10.1101/2020.04.21.20074732>.
- Niehus R, Martinez de Salazar Munoz P, Taylor A, Lipsitch M (2020) Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers. Preprint, submitted February 18, <https://www.medrxiv.org/content/10.1101/2020.02.13.20022707v2>.
- Nocedal J, Wright S (2006) *Numerical Optimization*. (Springer, New York).
- Perkins A, Espana G (2020) NotreDame-FRED COVID-19 forecasts. https://github.com/confunguido/covid19_ND_forecasting.
- PSI-DRAFT (2020) Disease rapid analysis and forecasting tool. Retrieved October 10, <https://zoltardata.com/model/254>.
- Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, Bracher J, et al (2020) Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. Preprint, submitted August 20,

- <https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1>.
- Rodriguez A, Tabassum A, Cui J, Xie J, Ho J, Agarwal P, Adhikari B, Prakash BA (2020) Deepcovid: An operational deep learning-driven framework for explainable real-time COVID-19 forecasting. Preprint, submitted March 21, <https://www.medrxiv.org/content/10.1101/2020.09.28.20203109v3>.
- Sayampanathan AA, Heng CS, Pin PH, Pang J, Leong TY, Lee VJ (2021) Infectivity of asymptomatic vs. symptomatic COVID-19. *Lancet*. 397(10269):93–94.
- Sood N, Simon P, Ebner P, Eichner D, Reynolds J, Bendavid E, Bhattacharya J (2020) Seroprevalence of Sars-CoV-2-specific antibodies among adults in Los Angeles County, California, on April 10–11, 2020. *JAMA*. Published online May 14, 2020, doi: 10.1001/jama.2020.8438.
- Streeck H, Schulte B, Kuemmerer B, Richter E, Höller T, Fuhrmann C, Bartok E, et al (2020) Infection fatality rate of SARS-CoV-2 infection in a german community with a super-spreading event. Preprint, submitted June 2, <https://www.medrxiv.org/content/10.1101/2020.05.04.20090076v2>.
- Wang C, Liu L, Hao X, Guo H, Wang Q, Huang J, He N, et al (2020) Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of coronavirus disease 2019 in Wuhan, China. Preprint, submitted March 6, <https://www.medrxiv.org/content/10.1101/2020.03.03.20030593v1>.
- Wise J (2020) Covid-19: Surveys indicate low infection level in community. *BMJ* 2020;369:m1992.
- Woody S, Tec MG, Dahan M, Gaither K, Lachmann M, Fox S, Meyers LA, Scott JG (2020) Projections for first-wave COVID-19 deaths across the us using social-distancing measures derived from mobile phones. Preprint, submitted April 26, <https://www.medrxiv.org/content/10.1101/2020.04.16.20068163v2>.
- Xiang Y, Sun D, Fan W, Gong X (1997) Generalized simulated annealing algorithm and its application to the thomson model. *Phys. Lett. A*. 233(3):216–220.
- Xu H, Huang S, Liu S, Deng J, Jiao B, Ai L, Xiao Y, Yan L, Li S (2020) Evaluation of the clinical characteristics of suspected or confirmed cases of COVID-19 during home care with isolation: A new retrospective analysis based on O2O. Preprint, submitted March 6, <https://ssrn.com/abstract=3548746>.
- Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D (2020) Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the diamond princess cruise ship: A data-driven analysis. *Int. J. Infect. Dis.* 93:201–204.

Michael Lingzhi Li is a graduating PhD candidate at the Massachusetts Institute of Technology Operations Research Center. His research interests include using data-driven analytics and optimization to create real-world impacts in healthcare and public policy.

Hamza Tazi Bouardi is current a senior data scientist at the Boston Consulting Group. He received a BS in Engineering and MS in Applied Mathematics from CentraleSupélec and a Masters in business analytics from Massachusetts Institute of Technology.

Omar Skali Lami is a graduating PhD candidate in Operations Research at the Massachusetts Institute of Technology. His research lies at the intersection of modern optimization with statistics and machine learning and their applications in healthcare, pricing and revenue management, and supply chain management.

Thomas Trikalinos is the Professor of Health Services, Policy and Practice and the Director of the Center for Evidence Synthesis in Health at Brown University. His research interests include optimizing the processes of evidence/synthesis by porting methodologies from computer science and applied mathematics, in particular decision making under deep uncertainty.

Nikolaos (Nikos) Trichakis is the Zenon Zannetos (1955) Career Development Professor and an associate professor of operations management at the MIT Sloan School of Management. His research interests include optimization under uncertainty, data-driven optimization, and analytics, with applications in healthcare, supply chain management, and finance.

Dimitris Bertsimas is the Boeing professor of operations research and associate dean of business analytics at Sloan School of Management, Massachusetts Institute of Technology. His research interests include machine learning and optimization and their applications in healthcare. He is currently editor in chief of *INFORMS Journal on Optimization*.