PNAS Nexus, 2024, 3, pgae204
https://doi.org/10.1093/pnasnexus/pgae204
Advance access publication 22 May 2024
Research Report

Reporting delays: A widely neglected impact factor in COVID-19 forecasts

Long Ma, Zhihao Qiu, Piet Van Mieghem and Maksim Kitsak*

Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, GA 2600, The Netherlands
*To whom correspondence should be addressed: Email: maksim.kitsak@gmail.com
Edited By: Rui Reis

Abstract

Epidemic forecasts are only as good as the accuracy of epidemic measurements. Is epidemic data, particularly COVID-19 epidemic data, clean, and devoid of noise? The complexity and variability inherent in data collection and reporting suggest otherwise. While we cannot evaluate the integrity of the COVID-19 epidemic data in a holistic fashion, we can assess the data for the presence of reporting delays. In our work, through the analysis of the first COVID-19 wave, we find substantial reporting delays in the published epidemic data. Motivated by the desire to enhance epidemic forecasts, we develop a statistical framework to detect, uncover, and remove reporting delays in the infectious, recovered, and deceased epidemic time series. Using our framework, we expose and analyze reporting delays in eight regions significantly affected by the first COVID-19 wave. Further, we demonstrate that removing reporting delays from epidemic data by using our statistical framework may decrease the error in epidemic forecasts. While our statistical framework can be used in combination with any epidemic forecast method that intakes infectious, recovered, and deceased data, to make a basic assessment, we employed the classical SIRD epidemic model. Our results indicate that the removal of reporting delays from the epidemic data may decrease the forecast error by up to 50%. We anticipate that our framework will be indispensable in the analysis of novel COVID-19 strains and other existing or novel infectious diseases.

Keywords: COVID-19 pandemic, reporting delays, epidemic forecasts, parametric optimization, SIRD compartmental model, time series analysis

Significance Statement

Accurate forecasts constitute the first step toward controlling an epidemic outbreak. However, the accuracy of such forecasts strongly depends on input data. Here, we develop a statistical framework to identify and remove reporting delays from the epidemic data. We use this framework to de-noise epidemic data of the first COVID-19 wave in eight regions worldwide. We demonstrate that the removal of reporting delays may increase the accuracy of epidemic forecasts by up to 50%. We anticipate that our framework will be invaluable in the analysis and forecasts of existing and future epidemics.

Introduction

The COVID-19 pandemic has been crippling the world's health, economies, and quality of life for over 3 years. We are gradually understanding that COVID-19 is here to stay. Fast mutation rates of the virus and its overwhelming spreading capability make it extremely hard, if not impossible, to eradicate (1). COVID-19 is not the first and almost surely not the last pandemic to hit humanity. Therefore, in order to better prepare and withstand other contagious diseases in the future, we need to extract as many lessons as possible from the COVID-19 pandemic.

Public awareness is, arguably, the first line of defense against any infectious disease. Efficient collection of epidemic data and accurate epidemic forecasts allow for timely containment of the spread or *flattening of the curve* to win time for the development of pharmaceutical treatment methods. Due to the success of network epidemiology and the broad availability of data, significant advances in epidemic modeling and forecast methods (2–6) have been achieved. However, the accuracy of epidemic forecasts strongly depends on the accuracy and timeliness of the input data.

Are epidemic data—especially in the short-time period after the onset of the epidemic—devoid of inaccuracies? Multiple sources suggest the negative answer. Indeed, public health indicators are known to be biased, because they are sensitive to fluctuations in supply and demand for diagnostic testing (7–9). Likewise, inequalities in geographic accessibility have been documented to adversely affect the coverage and timelines of health indicators (10, 11). It is now well understood that ignoring epidemic data inaccuracies may lead to delays in deploying intervention policies resulting in dire consequences of an epidemic on the population's health (12, 13).

One prominent challenge is the existence of temporal delays in data reporting (14, 15), defined as the time difference between the

Competing Interest: The authors declare no competing interest.
Received: July 24, 2023. **Accepted:** May 13, 2024

event when the person was affected by the virus and the time this event is accounted for. The reporting delays may occur due to many reasons, including patient hesitancy, medical testing delays, and reporting delays by public health authorities (16, 17). Consequently, reporting delays may vary not only across different diseases but also across different regions of interest. The median diagnosis delay for malaria is, for instance, approximately 4 days (18). Similarly, the research on the Middle East respiratory syndrome coronavirus (MERS-CoV) found a time difference between the symptom onset and confirmation of approximately four days (19). Hepatitis A, measles, and mumps data are usually reported after eight days (16). Reporting delays for some diseases are significantly longer: Hepatitis B, Shigellosis, and Salmonella can take on average 2–3 weeks (16). Recent COVID-19 studies reveal significant reporting delays of infections in China (20–25), Italy (26), Germany (27, 28), Singapore (29), the United States of America (30), and the United Kingdom (14).

While there is a plethora of scientific works aimed at the reconstruction of network data, including the reconstruction of network data for epidemic forecasts (31–33), we lack inference methods to assess and improve the accuracy of epidemic data itself. Instead, the prevalent epidemic forecast philosophy is focused either on data fusion (34–36)—or machine learning methods (37–39).

In our work, we develop a statistical framework to remove reporting delays. We demonstrate that the removal of reporting delays may significantly improve the accuracy of epidemic forecasts. Our statistical framework can be used in combination with any epidemic forecast method that takes infectious, recovered, and deceased epidemic data as input. Our work is organized as follows. We first present the evidence for the presence of reporting delays in the epidemic data extracted from the first COVID-19 wave. We then proceed to develop a statistical framework to remove reporting delays. After validating our framework on synthetic data, we move on to remove and analyze reporting delays in eight hotspots of the COVID-19 pandemic. We conclude our work with a discussion of the impact of delay removal on the accuracy of epidemic forecasts.

Results

Notation

Before presenting our findings, we introduce our notation for the epidemic data. Throughout the text, we operate with the infectious I, recovered R, and deceased D data. Each dataset is a time series of values, each corresponding to a specific observation time. For brevity, we refer to the triplet of infectious, recovered, and deceased data as $Y = \{I, R, D\}$. All values contained in the Y time series are fractions of individuals found in the corresponding state on a specific day. For instance, $I[k]$ corresponds to the fraction of individuals who are infectious on day k. Since COVID-19 reported data often is advertised in the form of changes in the number of epidemic cases, we find it convenient to introduce the daily changes in epidemic data as $\Delta Y[k]$ for $Y = \{I, R, D\}$. Further, in this work, we operate with reported epidemic data $\tilde{Y}$ and inferred data $\hat{Y}$. Since reported and inferred data are expected to differ from the true data Y, we need to distinguish the three. We summarize our notation in Table 1.

Table 1. Naming convention for the epidemic data, $Y = \{I, R, D\}$.

Cumulative quantities	Quantity increments
Y: fractions of cases	ΔY: fractions of new cases
$\tilde{Y}$: fractions of reported cases	$\Delta\tilde{Y}$: fractions of reported new cases
$\hat{Y}$: fractions of predicted cases	$\Delta\hat{Y}$: fractions of predicted new cases

Evidence for reporting delays in epidemiological data

We begin the exposition by considering daily reports on the fractions of infected $\Delta\tilde{I}$, recovered $\Delta\tilde{R}$, and deceased $\Delta\tilde{D}$ individuals in Spain. We used the daily epidemic reports in Spain to construct the fraction of infected individuals as $\tilde{I}[k] = \sum_{\ell=0}^{k-1} (\Delta\tilde{I}[\ell] - \Delta\tilde{R}[\ell] - \Delta\tilde{D}[\ell])$. While both the infected $\tilde{I}$ and the deceased data $\tilde{D}$ indicate that the first COVID-19 wave in Spain peaked in April 2020, the exact timings of the two peaks are, nevertheless, different. As observed in Fig. 1a, reported new deceased cases $\Delta\tilde{D}$ reached their peak on 2020 April 1st, while the highest fraction of infectious individuals $\tilde{I}$ was observed 22 days later on 2020 April 23. This observation is not specific to Spain: the peaks in the number of reported new deceased cases $\Delta\tilde{D}$ precede those of infectious cases by more than 1 week in most regions, Fig. S1. We make similar observations for COVID-19 daily recovery reports $\Delta\tilde{R}$, which exhibit their peaks after those of the deceased cases $\Delta\tilde{D}$, Figs. 1a and S1. Furthermore, when plotted as a function of daily deceased cases $\Delta\tilde{D}$, infectious cases $\tilde{I}$ and daily recovered cases $\Delta\tilde{R}$ form "loop" patterns, see Figs. 1b, c and S2.

The patterns observed in Figs. 1a–c, S1, and S2 may indicate the presence of reporting delays. Indeed, epidemic models view recoveries and deaths of patients as stochastic processes with effective rates proportional to the number of infected individuals I, $\Delta R[k] \propto \gamma_r I[k]$, $\Delta D[k] \propto \gamma_d I[k]$. Since the SARS-CoV-2 virus has hardly changed during the first wave of the pandemic, we expect that recovery γ_r and death γ_d rates are approximately constant during this period (40–42). This observation suggests that changes in the fractions of recovered ΔR, and deceased ΔD data are proportional to the fraction of infectious individuals I and, therefore, reach maximum values at the same time step, contradicting Fig. 1a–c.

We hypothesize that the disagreement between $\Delta\tilde{R}$, $\Delta\tilde{D}$, and $\tilde{I}$ is due to reporting delays. If each $\Delta\tilde{R}$, $\Delta\tilde{D}$, and $\tilde{I}$ time series are reported with different delays, their peaks are expected to differ. Likewise, the loop patterns of Fig. 1b, c may also be the result of an effective time shift between two nonmonotonous time series.

To check if reporting delays may result in the observed patterns, we consider the compartmental Susceptible-Infectious-Recovered-Deceased (SIRD) epidemic model. Within the SIRD model, the population is split into four compartments: susceptible S, infectious I, recovered R, and deceased D. Compartment S denotes the fraction of susceptible individuals, who can be infected by infectious individuals. Compartment I denotes the fraction of individuals, who have been infected but have not recovered or are deceased. Compartments R and D are respectively the fractions of individuals, who have recovered or are deceased. The SIRD model assumes that recovered individuals become immune and cannot be infected by the virus in the future. Further, the SIRD model assumes the uniform mixing of the Infectious and Susceptible sub-populations. As a result, the discrete-time transitions between the compartments are governed by first-order difference time equations

$$\begin{aligned} I[k+1] - I[k] &= \beta I[k] S[k] - (\gamma_r + \gamma_d) I[k], \\ R[k+1] - R[k] &= \gamma_r I[k], \\ D[k+1] - D[k] &= \gamma_d I[k], \\ S[k] + I[k] + R[k] + D[k] &= 1, \end{aligned} \tag{1}$$

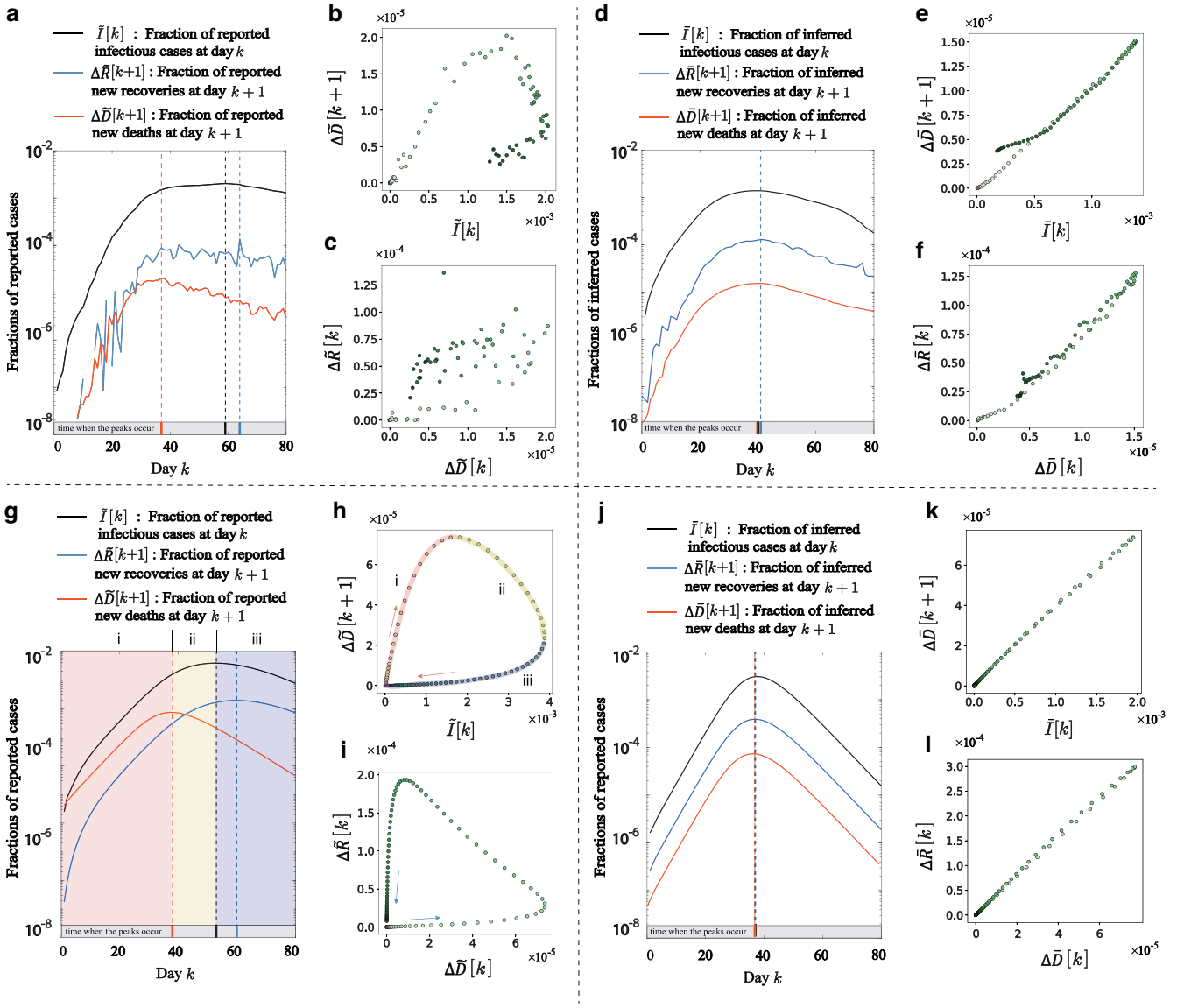

Fig. 1. Evidence for reporting delays in COVID-19 epidemic data. a) Reported infectious $\tilde{I}[k]$, recovered $\Delta\tilde{R}[k+1]$ and deceased $\Delta\tilde{D}[k+1]$ cases for the first COVID-19 wave in Spain. The $k=0$ day corresponds to 2020 February 25. Note the difference in peaks of the reported data. b) and c) display pairwise color-coded scatter plots of $\Delta\tilde{D}[k+1]$ vs. $\tilde{I}[k]$, and $\Delta\tilde{R}[k]$ vs. $\Delta\tilde{D}[k]$ for Spain. Colors, from light to dark green, reflect different days in the data ranging, respectively, from $k=0$ to $k=80$. The scatter plots in b) and c) form, respectively, clockwise and counterclockwise loop patterns. d)–f) display the epidemic data in Spain after the removal of reporting delays. g) Synthetic epidemic data generated with the SIRD model by solving Eq. 1 with parameters $\beta=0.5$, $\gamma_r=0.2$, $\gamma_d=0.05$, $R[0]=D[0]=0$, $I[0]=10^{-6}$, and $S[0]=1-I[0]$. Synthetic reporting delays were generated with the Pólya-Aeppli distributions using Eq. 3 with parameters $\lambda_I=3.68$, $\theta_I=0.5$, $E[T_I]=7.36$; $\lambda_R=9.46$, $\theta_R=0.4$, $E[T_R]=23.65$; $\lambda_D=0.3$, and $\theta_D=0.3$, $E[T_D]=1$. h) and i) display pairwise color-coded scatter plots of $\Delta\tilde{D}[k+1]$ vs. $\tilde{I}[k]$, and $\Delta\tilde{R}[k]$ vs. $\Delta\tilde{D}[k]$ for the SIRD epidemic data. j)–l) display the synthetic epidemic data after the removal of reporting delays.

where β, γ_r, and γ_d are the infection, the recovery, and the deceased probabilities, respectively. We used Eq. 1 to generate infectious $I[k]$, recovered $R[k]$, and deceased $D[k]$ time series epidemic data, and then added synthetic delays to the time series data, see Fig. 1g–i and Supplementary Material B. Since reporting delays were manually added, we observe that $\Delta R[k+1]$, $\Delta D[k+1]$, and $I[k]$ reach maxima at different time steps, Fig. 1g.

Upholding our hypothesis, we observe the formation of loop patterns in the $\Delta\tilde{D}[k+1]$ vs. $\tilde{I}[k]$ and $\Delta\tilde{R}[k+1]$ vs. $\tilde{D}[k]$ in Fig. 1h, i, similar to those of Spain, Fig. 1b, c. The observed loop patterns are due to the effective horizontal shifts of the corresponding times series due to reporting delays. Indeed, let us split the observation time window into three windows formed by the maxima of the $\Delta\tilde{D}[k+1]$ and $\tilde{I}[k]$ curves, as shown in Fig. 1g. In window i, both $\tilde{I}[k]$ and $\Delta\tilde{D}[k+1]$ increase as a function of discrete time k. Since reporting delays in the synthetic $\Delta\tilde{D}[k+1]$ data are smaller than those in $\tilde{I}[k]$, this time window corresponds to the upper branch of the $\Delta\tilde{D}[k+1]$ vs. $\tilde{I}[k]$ loop in Fig. 1h. In window ii, $\tilde{I}[k]$ increases while $\Delta\tilde{D}[k+1]$ decreases. Thus, window ii corresponds to the top (decreasing) section of the $\Delta\tilde{D}[k+1]$ vs. $\tilde{I}[k]$ loop, Fig. 1h. Finally, in window iii both $\Delta\tilde{D}[k+1]$ and $\tilde{I}[k]$ decrease as a function of time step k resulting in the lowest section of the $\Delta\tilde{D}[k+1]$ vs. $\tilde{I}[k]$ loop, Fig. 1h. Combined, all sections correspond to the loop pattern of Fig. 1h with points progressing in the clockwise direction. Similar considerations explain the counterclockwise loop pattern in the $\Delta\tilde{R}[k+1]$ vs. $\tilde{D}[k]$ scatter plot. The counterclockwise progression of points in this loop pattern is due to $\Delta\tilde{R}[k]$ lagging behind the $\Delta\tilde{D}[k]$ time series, Fig. 1i.

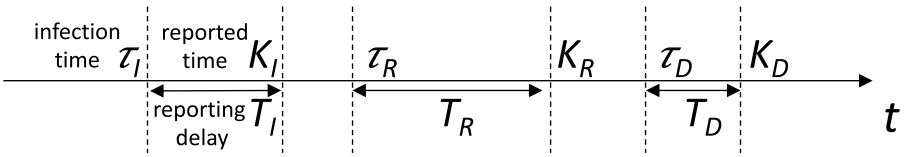

Fig. 2. A schematic representation of epidemic events and corresponding delay times.

A null model for reporting delays

To uncover reporting delays, we employ the following null model: each individual i is endowed with random event time τ_{Y_i} of either getting infected, τ_{I_i}, recovered, τ_{R_i}, or deceased, τ_{D_i}. Figure 2 depicts the random delay time for event $Y_i = \{I_i, R_i, D_i\}$ as $T_{Y_i} = \{T_{I_i}, T_{R_i}, T_{D_i}\}$ and the corresponding random reported time as $K_{Y_i} = \tau_{Y_i} + T_{Y_i}$. Within our discrete-time setting, these random times are discrete random variables. Assuming that the reporting delays T_{Y_i} are independent of the events Y_i, we obtain for the reported time K_{Y_i}

$$\Pr[K_{Y_i} = k] = \sum_{y=0}^{\infty}\sum_{m=0}^{\infty} \Pr[\tau_{Y_i} = y]\Pr[T_{Y_i} = m]\delta_{y+m,k}$$
$$= \sum_{m=0}^{k} \Pr[\tau_{Y_i} = k - m]\Pr[T_Y = m]. \quad (2)$$

By taking the arithmetic mean of Eq. 2, we find that the expected reported fraction $\Delta\widetilde{Y}[k] = \frac{1}{N}\sum_i \Pr[K_{Y_i} = k]$ of individuals in state Y is a discrete convolution

$$\Delta\widetilde{Y}[k] = \sum_{m=0}^{k} \Pr[T_Y = m]\Delta Y[k - m], \quad (3)$$

where $Y = \{I, R, D\}$. Equation 3 serves as the foundation for uncovering reporting delays and improving epidemic forecasts.

Statistical framework to uncover reporting delays

The delay distribution $\Pr[T_Y = m]$ in (3) can be determined, in principle, as the solution of a large set of quadratic equations, deduced from the governing epidemic equations, as shown in Supplementary Material D, provided that we assume a same distribution for each delay $T_{Y_i} = \{T_{I_i}, T_{R_i}, T_{D_i}\}$. Since that solution is rather demanding, we propose a simplification. We assume that the delay distribution $\Pr[T_Y = m]$ is generated by a two-parameter probability distribution with parameters $\kappa = \{\lambda, \theta\}$, which is sufficiently general to incorporate the real probability distribution of the delay.

We consider three families of two-parameter discrete probability distributions: the negative binomial distribution, the Neyman type A distribution, and the Pólya-Aeppli distribution, see Materials and methods A. These distributions are quite general and contain some well-known 1-parameter distributions as special cases. The logarithmic, geometric, and Poisson distributions are, for instance, all special cases of the negative binomial distribution (43). The three distributions are sufficiently general to generate data with variable mean and variance and of varying skewness (43, 44)—properties expected of the epidemic delay data, see Materials and methods A. For brevity, we focus on the Pólya-Aeppli distribution in the main text and report the results for the other two distributions in Fig. S4.

In our statistical framework, we assume that the reporting delays correspond to three datasets, $Y = \{I, R, D\}$, which are all characterized by the Pólya-Aeppli distribution, albeit with different parameters $\kappa_Y \equiv \{\lambda_Y, \theta_Y\}$. Hence, there are in total six parameters for three Pólya-Aeppli distributions, $\boldsymbol{\kappa} = (\lambda_I, \theta_I, \lambda_R, \theta_R, \lambda_D, \theta_D)$. With the choice of the Pólya-Aeppli distribution, the reporting delays $\Delta\widetilde{Y}$ in our basic equation Eq. 3 can be determined in the parameterized form $\Delta\widetilde{Y}_{\boldsymbol{\kappa}}$.

Given the incremental time series $\Delta\widetilde{Y}_{\boldsymbol{\kappa}}$, the cumulative time series $\widetilde{Y}_{\boldsymbol{\kappa}}$ are given by

$$\begin{aligned}\widetilde{I}_{\boldsymbol{\kappa}}[k+1] &= \widetilde{I}_{\boldsymbol{\kappa}}[k] + \Delta\widetilde{I}_{\boldsymbol{\kappa}}[k] - \Delta\widetilde{R}_{\boldsymbol{\kappa}}[k] - \Delta\widetilde{D}_{\boldsymbol{\kappa}}[k],\\ \widetilde{R}_{\boldsymbol{\kappa}}[k+1] &= \widetilde{R}_{\boldsymbol{\kappa}}[k] + \Delta\widetilde{R}_{\boldsymbol{\kappa}}[k],\\ \widetilde{D}_{\boldsymbol{\kappa}}[k+1] &= \widetilde{D}_{\boldsymbol{\kappa}}[k] + \Delta\widetilde{D}_{\boldsymbol{\kappa}}[k],\end{aligned} \quad (4)$$

for $k \geq 0$, and $\widetilde{I}_{\boldsymbol{\kappa}}[0] = \widetilde{R}_{\boldsymbol{\kappa}}[0] = \widetilde{D}_{\boldsymbol{\kappa}}[0] = 0$.

The main assumption of our reporting delay removal framework is that the increments in new recovered ΔR and deceased ΔD individuals are proportional to the cumulative fraction of infectious individuals I.

Therefore, we determine the "best" parameters $\bar{\boldsymbol{\kappa}}$ that maximize the product of pairwise correlations among the three epidemic time series in $\widetilde{Y}_{\boldsymbol{\kappa}}$:

$$O_b(\widetilde{Y}_{\boldsymbol{\kappa}}) \equiv O_b(\widetilde{I}_{\boldsymbol{\kappa}}, \Delta\widetilde{R}_{\boldsymbol{\kappa}}, \Delta\widetilde{D}_{\boldsymbol{\kappa}}) = \rho(\Delta\widetilde{R}_{\boldsymbol{\kappa}}, \Delta\widetilde{D}_{\boldsymbol{\kappa}})\rho(\widetilde{I}_{\boldsymbol{\kappa}}, \Delta\widetilde{R}_{\boldsymbol{\kappa}})\rho(\widetilde{I}_{\boldsymbol{\kappa}}, \Delta\widetilde{D}_{\boldsymbol{\kappa}}), \quad (5)$$

where $\rho(X, Y)$ is the Pearson correlation coefficient between time series X and Y. The objective function $O_b(\widetilde{Y}_{\boldsymbol{\kappa}})$ in Eq. 5 reaches its maximum value of 1 when all three pairwise correlations among the $\widetilde{I}_{\boldsymbol{\kappa}}$, $\Delta\widetilde{R}_{\boldsymbol{\kappa}}$ and $\Delta\widetilde{D}_{\boldsymbol{\kappa}}$ time series are 1, which we expect when recovery γ_r and deceased γ_d epidemic probabilities are constant. Due to the nature of the Pearson correlation coefficient, the objective function $O_b(\widetilde{Y}_{\boldsymbol{\kappa}}[k]) = O_b(\widetilde{Y}_{\boldsymbol{\kappa}}[k - T])$ is invariant under the constant time shift T of the epidemic data. As a result, we can only infer the reporting delays up to a constant time shift T.

To maximize $O_b(\widetilde{Y}_{\boldsymbol{\kappa}})$, the random search (45) method is applied: we conduct a large set of independent random iterations, $\ell = 1, \ldots, L$. At each iteration ℓ, we select the elements of the parameter vector $\bar{\kappa}$ uniformly at random from the prescribed domain of values. Further, at each iteration ℓ, we use the selected parameter set $\bar{\boldsymbol{\kappa}}_\ell$ to reconstruct the original epidemic data $\Delta\widetilde{Y}_{\bar{\boldsymbol{\kappa}}_\ell}$ by solving Eq. 3 and convert incremental data $\Delta\widetilde{Y}_{\bar{\boldsymbol{\kappa}}_\ell}$ to cumulative data $\widetilde{Y}_{\bar{\boldsymbol{\kappa}}_\ell}$ by solving Eq. 4. We then compute the pairwise correlations in $\widetilde{Y}_{\bar{\boldsymbol{\kappa}}_\ell}$ to obtain the objective function $O_b(\widetilde{Y}_{\bar{\boldsymbol{\kappa}}_\ell})$. After completing all random search iterations, the resulting delay parameter $\hat{\boldsymbol{\kappa}}$ is the one maximizing the objective function, $\hat{\boldsymbol{\kappa}} = \mathrm{argmax}_{\bar{\boldsymbol{\kappa}}_\ell} O_b(\widetilde{Y}_{\bar{\boldsymbol{\kappa}}_\ell})$, see Materials and methods B.

Tests on synthetic data

Before uncovering reporting delays in real epidemic data, we test our framework on synthetic datasets, which we generate with the SIRD compartmental model, Eq. 1.

To test our framework, we generated 50 SIRD epidemic model datasets with different epidemic parameters and added synthetic reported delays to the obtained times series, as prescribed by Eq. 3. The resulting delay-perturbed synthetic data is fed to our statistical framework that infers the reported synthetically added delays. Figure 1d–f and j–l display, respectively, the reconstructed

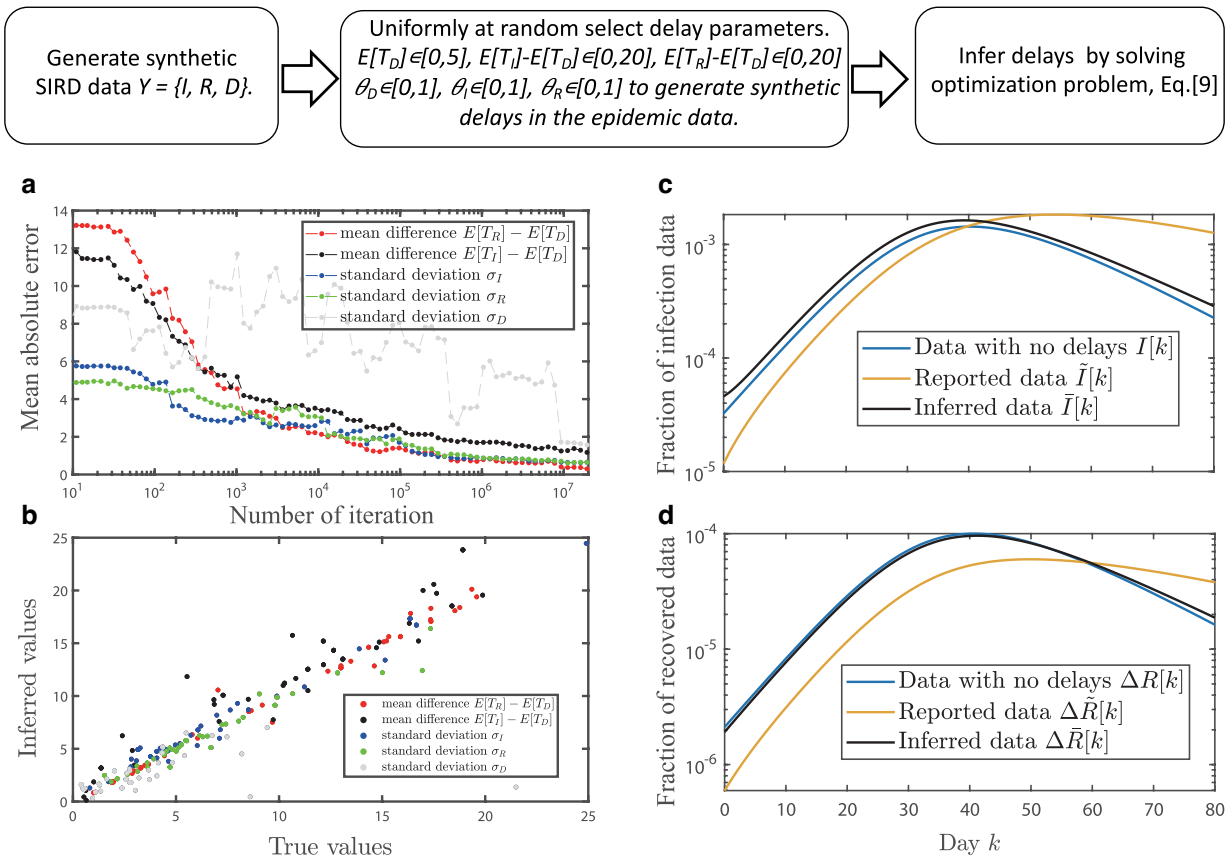

Fig. 3. Uncovering reporting delays in synthetic data. We generate epidemic data using the SIRD model, Eq. 1 with parameters $\beta = 0.23$, $\gamma_r = 0.079$, and $\gamma_d = 0.11$ and variable delay parameters. The schematics on top illustrates the process of generation and inference on synthetic data. a) The mean absolute inference errors as functions of the number of iteration steps in the random search. Due to the nature of the Pearson correlation coefficient, it is only possible to infer relative delays, $E[T_R] - E[T_D]$, $E[T_I] - E[T_D]$, and $E[T_R] - E[T_I]$. b) The relationship between the true and the inferred values of delay parameters. c, d) An example of an SIRD synthetic data c) before and d) after the removal of synthetic reporting delays.

real and synthetic epidemic data. We find strong correlations between the reconstructed $\Delta\tilde{R}$, $\Delta\tilde{D}$, and $\tilde{I}$ data. Further, Fig. 3a, b indicate that the inferred delay parameters are in good agreement with the true parameters that generated the datasets. Figure 3 illustrates that the inference errors decrease fast as a function of the number of iteration steps saturating at the value of 2 days after 10^7 iterations. To assess the robustness of the inference procedure, we have conducted cross-inference experiments by generating synthetic data with one delay distribution and inferring delays using another distribution, arriving at similar results in Fig. S4.

Uncovering reporting delays in real data

After testing our inference framework on synthetic data, we moved on to uncover reporting delays in real epidemic data that we have collected from eight regions worldwide, Supplementary Material A. Figure 1d–f and j–l display the reconstructed epidemic data for Spain and for synthetic SIRD model after the identification and removal of reporting delays, see Table S1 for the inferred parameters. Consistent with our expectations, the reconstructed time series of infected $I[k]$, new recovered $\Delta R[k+1]$, and new deceased $\Delta D[k+1]$ individuals are strongly correlated, Fig. 1e, f and k, l, and their peaks co-occur, as seen in Fig. 1d, j.

We summarize the properties of uncovered reporting delays in the eight regions in Table 2. We find that the infectious and recovered data delays are longer than those for deceased cases and vary from several days to several weeks. This observation is hardly surprising. Indeed, there are several factors contributing to the delays of infectious cases. One factor is the delay between an individual becoming infectious, and the symptom onset (46). Another factor, particularly significant in the first COVID-19 wave, is the delay between the symptom onset and the test result (47). In their turn, delays in the recovered events are likely caused by hospital discharge policies. Indeed, the recovered data are usually derived from hospital discharge events, which occur after the patient's recovery.

Based on the expected delay values, one can naturally split the eight ROIs into three categories: (i) large infectious and small recovered delays: Romania, Germany, and Denmark, (ii) small infectious and large recovered delays: Italy and Spain, and (iii) large infectious and recovered delays: Wuhan, Hubei, and Turkey.

Small infectious delays imply that COVID-19 tests are timely and accurate. This seems to be the case for Italy, which executed more tests per capita in April 2020 than other countries (48). On the contrary, the testing ability of the Hubei province was significantly insufficient during the first wave of the COVID-19 outbreak.

Large delays in the recovered data may be attributed to strict hospital discharge policies. Discharge policy in Italy, for instance, was based on the negative test (49), and it has been shown that COVID-19 tests may stay positive for an extended time after COVID-19 symptoms disappear. In contrast, discharge policies

Table 2. Inferred reporting delays.

Regions	$E[T_R] - E[T_D]$	$E[T_I] - E[T_D]$	σ_I	σ_R	σ_D	$\rho(\Delta\bar{R}, \Delta\bar{D})$	$\rho(\bar{I}, \Delta\bar{R})$	$\rho(\bar{I}, \Delta\bar{D})$	$O_b(\bar{Y})$
Italy	28.52	6.59	5.08	27.79	0.72	0.91	0.94	0.99	0.84
Spain	22.66	6.36	8.63	28.39	0.73	0.99	0.99	1.00	0.98
Wuhan	14.80	20.64	51.67	16.13	23.87	0.78	0.89	0.91	0.63
Turkey	14.13	24.85	43.89	12.47	11.21	0.91	0.95	0.98	0.85
Hubei	9.96	21.07	78.62	10.89	54.14	0.85	0.90	0.92	0.71
Romania	3.30	19.05	43.08	90.12	0.29	0.81	0.89	0.97	0.70
Germany	2.72	16.55	39.25	106.89	4.31	0.87	0.88	0.99	0.76
Denmark	0.17	25.07	36.90	2.71	28.25	0.83	0.93	0.94	0.72

The table displays the inferred differences between the expected delay times $E[T_R] - E[T_D]$, $E[T_I] - E[T_D]$, standard deviations σ_I, σ_R and σ_D, and optimized values for the objective function $O_b(\bar{Y})$. See Tables S1 and S2 for the corresponding parameters of the Pólya-Aeppli distributions.

in Denmark were based not on the negative test but on patient symptoms (49), likely leading to shorter reporting delays in the recovered data.

Large standard deviations in the reporting delays may indicate irregularities in reporting mechanisms. As an example, Hubei province expanded its daily testing capacity from 200 to 2,000 individuals from the beginning of the pandemic until January 27th (50). As a result, fewer individuals were tested late at the end of the first COVID-19 wave compared to its beginning, likely resulting in the large standard deviation of infectious data delays.

The small standard deviation in the delays of recovered data observed in Hubei and its capital Wuhan is likely the consequence of the strict discharge criteria (51). Although strict discharge policies cause significant delays, these delays are similar, resulting in relatively smaller σ_R values. In contrast to Chinese regions, the recovery data for Germany are not reported directly but instead are estimated by a not explained algorithm (52), resulting in large errors and, consequently, larger σ_R values. Based on the optimized values $O_b(\bar{Y})$ as shown in Table S2, the optimization performances of our algorithm for Italy, Spain, and Turkey are better than the other countries.

Improving epidemic forecasts

Is accounting for reporting delays likely to improve epidemic forecasts? To answer this question, we designed two experiments. Experiment 1 aims to forecast the epidemics ignoring reporting delays and serves as a baseline for Experiment 2, which uncovers reporting delays prior to forecasting epidemic data.

In both experiments, we split the reported epidemic data into two parts, which we call the training and the testing sets, respectively, see Fig. 4a. In Experiment 1, we fit the training set with the SIRD epidemic model, obtaining model parameters β, $\gamma_r + \gamma_d$, and the fraction of initial infected cases $I[0]$. We then use the SIRD model with the obtained parameters to forecast the epidemic data, Supplementary Material C.

In Experiment 2, we first use the reported data $\widetilde{Y}$ in the training set to infer the parameters of the delay distributions $\bar{\boldsymbol{\kappa}}$. We rely on these parameters to remove reporting delays from the testing set and obtain reconstructed data $\Delta\bar{I}$. In the next step, we fit the reconstructed data $\Delta\bar{I}$ with the SIRD model, obtaining β, $\gamma_r + \gamma_d$ and $I[0]$ spreading parameters. We use these spreading parameters to forecast the epidemic data $\Delta\hat{I}$, which we compare to the reported data $\Delta\widetilde{I}$ in the testing set. Since $\Delta\widetilde{I}$ in the testing set contains the reporting delays, while the forecast $\Delta\hat{I}$ does not, we added the inferred reporting delays back to the $\Delta\hat{I}$ using Eq. 3, obtaining $\Delta\hat{I}^*$, See Fig. 4a and Supplementary Material C.

Figure 4b, c presents the results of the forecast experiments for Spain, indicating that correcting for reporting delays does improve the forecast accuracy. To evaluate the forecast errors in a systematic way, we evaluated the root mean square errors between the forecasts I_F and the testing I_T sets:

$$\mathrm{RMSE}(I_F, I_T) = \left[\frac{1}{n}\sum_{k=0}^{n-1}(I_F[k] - I_T[k])^2\right]^{\frac{1}{2}}, \tag{6}$$

where n is the size of the testing set. Further, to quantify the benefits of accounting for reporting delays, we used the ratio of the root mean square errors measured in forecasts with and without reporting delays, $\epsilon \equiv \frac{\mathrm{RMSE}(\hat{I}_2^*, I)}{\mathrm{RMSE}(\hat{I}_1, I)}$, where $\hat{I}_1$ and $\hat{I}_2^*$ are the epidemic forecasts obtained in experiments 1 and 2, respectively. The smaller the ratio, the smaller the relative forecast error.

We measured the accuracy of epidemic forecasts with different start dates for Spain in Germany. While we did observe a substantial improvement in forecast accuracy for Spain, Fig. 4d, this was not the case for Germany, where the removal of reporting delays only improved the forecasts by a small margin, Fig. 4e. On a broader scale, we observed that the benefits of correction for reporting delays vary across all regions of interest, Figs. 4f, S5, and S6. While there is nearly a two-fold improvement in the forecast accuracy for Denmark, there is little improvement for Wuhan and Hubei.

Apart from only measuring and testing a small part of the population—which might be too small or not sampled adequately—there are at least two factors that may hinder epidemic forecast accuracy. The first factor is the insufficient accuracy of reporting delay removals. The second one is the inability of the SIRD model to reproduce the COVID-19 dynamics accurately. We can quantify the former by the maximum attained value of the objective function $O_b(\bar{Y})_{\max}$ that we aim to maximize when removing reporting delays. While there is no direct way to quantify the goodness of the SIRD model in epidemic forecasts, as an indirect measure, we fit the training set with the SIRD model and compute the fitting error.

Figure 4g shows that the forecast error is strongly correlated with the SIRD fitting error, Pearson $r = 0.89$, $p < 10^{-27}$, indicating that the highest epidemic forecast accuracy is attained when the model fit is accurate. The forecast error ratio depends both on the $O_b(\bar{Y})_{\max}$ and SIRD model fitting, Fig. 4h. We observe the largest error ratio ϵ in Wuhan, Hubei, and Romania, Fig. 4h. Wuhan and Hubei correspond to the largest SIRD fitting errors, while Romania corresponds to the lowest SIRD fitting errors. At the same time, all three regions are quantified by the lowest $O_b(\bar{Y})_{\max}$ values. The other five regions, corresponding to smaller error ratios, are characterized by significantly larger $O_b(\bar{Y})_{\max}$ values, Fig. 4g and Table S2.

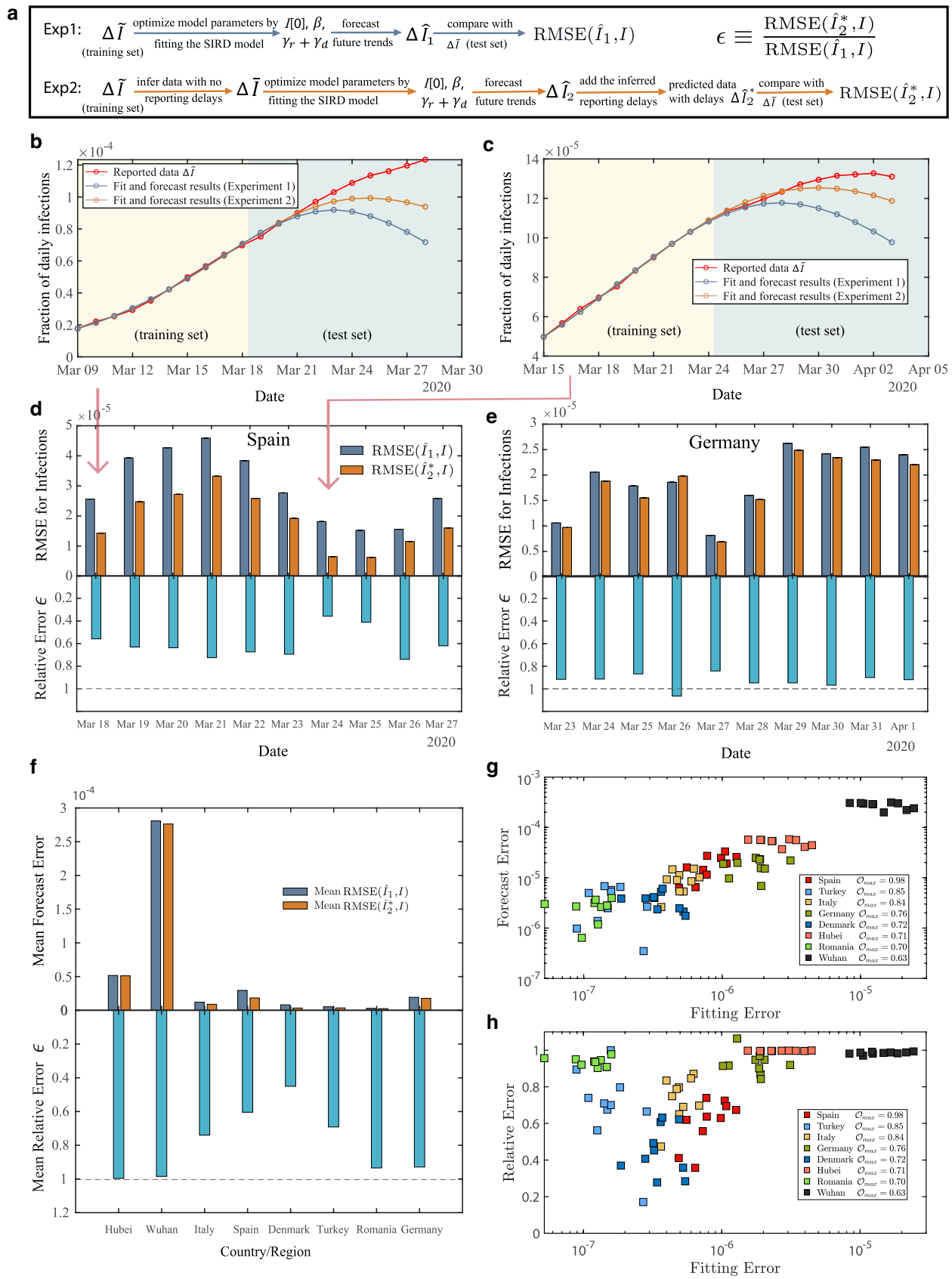

Fig. 4. Accounting for reporting delays improves epidemic forecasts. a) The schematic diagram for the two forecast experiments. b), c) display the results of epidemic forecasts in Spain with different forecast start dates. The forecasts of experiment 2 (yellow) are closer to the reported data (red) than the forecasts of experiment 1 (blue). The forecast results for all eight countries or regions are shown in Figs. S5 and S6. d), e) All forecast results for Spain and Germany. Shown are the RMSE forecast errors and their ratios. h) Average forecast errors ϵ for the 8 regions. The benefits of removing reporting delays vary by region. g) The average RMSE forecast error of experiment 2 as a function of SIRD model fitting error for all experiments. h) The mean relative forecast error as a function of SIRD model fitting error for all experiments.

These observations indicate that $O_b(\bar{Y})_{\max}$ may be used as an early indicator of the benefit of delay removals. Indeed, lower $O_b(\bar{Y})_{\max}$ values may indicate that reporting delays were not removed successfully or the initial assumption of the proportionality among I, ΔR, and ΔD times series does not hold. We note that high $O_b(\bar{Y})_{\max}$ values correspond to high benefits of delay removals regardless of the SIRD fitting error. In the case of Romania, for instance, SIRD fitting errors are among the smallest, resulting in low forecast errors. Yet, the removal of reporting delays does not result in even lower forecast errors.

There might be multiple reasons for suboptimal reporting delay removals in the case of Hubei, Wuhan, and Romania. One possibility is the nonstationary nature of delays. The main assumption of our framework is that delay times are independent and drawn from the same distribution. While COVID-19 did not significantly mutate over the short time span of the first wave, our ability to handle the infections improved significantly. The PCR test capacity in China has grown remarkably during the first COVID wave, possibly explaining the limited effect of reporting delays on improving epidemic forecasts in Hubei and Wuhan.

Discussion

Data delays are ubiquitous in data sciences and adversely affect data analysis (53). The unique property of the epidemic data, enabling us to identify and filter out reporting delays, is the proportionality between the number of infectious individuals I and the rates of change in the deceased ΔD and the recovered ΔR individuals.

We relied on this proportionality property to develop a parametric statistical framework to uncover reporting delays in the first COVID-19 wave and applied it to eight regions significantly affected during the first COVID-19 wave. The character of uncovered delays varies across studied regions and can be explained by region-specific medical capacities and epidemic policies.

Concerning the epidemic forecasts, we found that the benefits of using curated epidemic data, as opposed to the raw data, are maximized in situations when reporting delays are removed efficiently. One of the main factors hindering the efficiency of delay removal —the nonstationarity of reporting delays—can be caused by either varying spreading properties of the virus or by rapid changes in medical capacities or epidemic restrictions across the regions. While the spreading properties of COVID-19 did not change significantly during the first wave of the pandemic, both medical capacities and epidemic policies were the subjects of constant updates as the communities learned how to handle the pandemic.

We expect that our framework will prove most useful in the early stages of an epidemic when accurate epidemic forecasts are essential to plan intervention strategies and raise public awareness. Common sense dictates that data collection and reporting challenges are the most significant during that time.

Our statistical framework is based on the assumption that transition probabilities γ_r and γ_d quantifying transitions from the infectious to the recovered, $I \to R$ and the infections to the deceased, $I \to D$ states, respectively are constant across the population and do not change over the observational period. While these assumptions are approximations, we expect them to hold better in early epidemic stages than in later stages. In the later stages, the population becomes more heterogeneous, e.g. due to natural or vaccine-induced immunity (54). In these situations, our framework can be generalized by splitting the infectious compartment I into sub-compartments and considering different rates, e.g.

$$\begin{aligned} \Delta R[k] &= \gamma_{1,r} I_1[k] + \gamma_{2,r} I_2[k], \\ \Delta D[k] &= \gamma_{1,d} I_1[k] + \gamma_{2,d} I_2[k], \\ I[k] &= I_1[k] + I_2[k], \end{aligned} \tag{7}$$

where $I_1[k]$ and $I_2[k]$ are two infectious sub-compartments characterized by different recovery and death probabilities.

At the same time, we stress that our statistical framework does not make strong assumptions about the spreading mechanisms of a pathogen and can be used in combination with any forecasting method that requires infectious, recovered, and deceased data as input.

In conclusion, our framework can be used as a preliminary filter for any epidemic forecast tool that takes infected, recovered, and deceased data as input. Our framework holds for any compartmental epidemic model, provided that each compartment can be measured, because a time series of each compartment is indispensable. Accurate and timely epidemic forecasts are of immense value for society and policymakers to minimize the adverse effects of the virus.

Materials and methods

Types of distributions

To determine the family of reporting delay distributions that best suit our data, we consider three different two-parameter discrete distributions (55) below:

(I) Negative binomial distribution.

$$\Pr[T = m] = \binom{m + r - 1}{m} (1 - p)^m p^r. \tag{8}$$

The negative binomial distribution with parameters $r > 0$ and $p \in [0, 1]$ has mean value $E[T] = r(1 - p)/p$ and variance $\mathrm{Var}[T] = r(1 - p)/p^2$.

(II) Pólya-Aeppli distribution is also known as the geometric Poisson distribution.

$$\Pr[T = m] = \begin{cases} \sum_{j=1}^{m} e^{-\lambda} \frac{\lambda^j}{j!} (1 - \theta)^{m-j} \theta^j \binom{m-1}{j-1}, & m > 0 \\ e^{-\lambda}, & m = 0 \end{cases}. \tag{9}$$

The Pólya-Aeppli distribution with parameters $\lambda > 0$ and $\theta \in [0, 1]$ has mean value $E[T] = \lambda/\theta$ and variance $\mathrm{Var}[T] = \lambda(2 - \theta)/\theta^2$.

(III) Neyman type A distribution.

$$\Pr[T = m] = \frac{\mu^m e^{-\xi}}{m!} \sum_{j=0}^{\infty} \frac{(\xi e^{-\mu})^j}{j!} j^m. \tag{10}$$

The Neyman type A distribution with parameters $\xi > 0$ and $\mu > 0$ has mean value $E[T] = \xi\mu$ and variance $Var[T] = \xi\mu(1 + \mu)$.

Inferring reporting delays

To uncover reporting delays in epidemic data, we determine parameters of delays distributions $\boldsymbol{\kappa}_Y$ maximizing the pairwise correlations between the epidemic time series, Eq. 5. This optimization problem can be compactly written as:

$$\begin{aligned} \arg\max_{\boldsymbol{\kappa}} \quad & O_b(Y) \equiv \rho(\Delta R, \Delta D)\rho(I, \Delta R)\rho(I, \Delta D) \\ \text{s.t.} \quad & \Delta I = \Psi_I^{-1} \Delta \widetilde{I},\ \Delta R = \Psi_R^{-1} \Delta \widetilde{R},\ \Delta D = \Psi_D^{-1} \Delta \widetilde{D}, \\ & \min(\Delta I[k], \Delta R[k], \Delta D[k], I[k]) \geq 0, \quad \text{for} \quad k = 1, \ldots, T. \end{aligned} \tag{11}$$

Here T is the size of epidemic time series, $\widetilde{Y} \equiv \{\widetilde{I}, \widetilde{R}, \widetilde{D}\}$ and $Y \equiv \{I, R, D\}$ are reported and reconstructed epidemic data,

respectively, while $\Delta Y = \Psi_Y^{-1}\Delta\widetilde{Y}$ are the matrix solutions of Eq. 3. Indeed, Eq. 3 can be written in the matrix form as $\Delta\widetilde{Y} = \Psi_Y \Delta Y$ for $Y = \{I, R, D\}$, where

$$\Psi_{Y_{i,j}} \triangleq \begin{cases} \Pr[T_Y = i - j] & \text{if} \quad i \geq j. \\ 0 & \text{otherwise.} \end{cases}'$$

Then, Ψ^{-1} is the inverse of Ψ_Y and $\Delta Y = \Psi_Y^{-1}\Delta\widetilde{Y}$.

In the main text, we assume that reporting delays are iid random variables drawn from three Pólya-Aeppli distributions, Eq. 9, with distinct parameters $\{\lambda_I, \theta_I\}$, $\{\lambda_R, \theta_R\}$, and $\{\lambda_D, \theta_D\}$. In the case of Pólya-Aeppli distributions, the parameter vector takes the form of $\boldsymbol{\kappa} \equiv \{\lambda_I, \theta_I, \lambda_R, \theta_R, \lambda_D, \theta_D\}$. We solve the optimization problem given by Eq. 11 and the random search (45). For each iteration $\ell = 1, \ldots, L$, we treat the expected values of the Pólya-Aeppli distributions, $E[T_Y] = \lambda_Y/\theta_Y$, as iid random variables and draw from the uniform pdfs $U[0, 30]$ for $Y = \{I, R, D\}$. Similarly, we draw θ_Y parameters independently at random from uniform pdfs $U[0, 1]$ and then determine λ_Y parameters as $\lambda_Y = \theta_Y E[T_Y]$, obtaining $\boldsymbol{\kappa}_\ell$. In our experiments, we set the maximum number of search iterations to $L = 10^7$.

For each $\boldsymbol{\kappa}_\ell$, we use Eq. 3 to reconstruct original data Y_ℓ, which we then use to compute the objective function $O_b(Y_\ell)$. After completing all iterations, the thought parameter vector $\hat{\boldsymbol{\kappa}}$ describing reporting delays is the one corresponding to the maximum $O_b(Y_\ell)$ value, $\hat{\boldsymbol{\kappa}} = \arg\max_{\boldsymbol{\kappa}} \quad O_b(Y)$.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This research has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101019718). L.M. is thankful for the support from the China Scholarship Council. M.K. has been supported by the Dutch Research Council (NWO) grant OCENW.M20.244.

Author Contributions

L.M. conceived the research idea. L.M., P.V.M., and M.K. designed research. L.M. and Z.Q. performed research. All authors discussed results and wrote the manuscript.

This manuscript has been posted as a preprint in the ArXiv repository https://arxiv.org/abs/2304.11863.

Data Availability

All epidemic data used in this work are publicly available from the original sources, see Supplementary Information A. The extracted epidemic time series for the eight regions of interest are deposited in FigShare (https://doi.org/10.6084/m9.figshare.22639519.v1). Code to generate SIRD epidemic time series and to uncover reporting delays is available on GitHub (https://github.com/qzhszl/Reporting-delays-a-widely-neglected-impact-factor-in-COVID-19-forecasts.git).

References

1 Phillips N. 2021. The coronavirus is here to stay-here's what that means. *Nature*. 590(7846):382–384.
2 Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. 2015. Epidemic processes in complex networks. *Rev Mod Phys*. 87(3):925–979.
3 Kiss IZ, Miller JC, Simon PL. 2017. *Mathematics of epidemics on networks*. Vol. 598. Cham: Springer.
4 Brauer F. 2017. Mathematical epidemiology: past, present, and future. *Infect Dis Model*. 2(2):113–127.
5 Achterberg MA, *et al*. 2022. Comparing the accuracy of several network-based COVID-19 prediction algorithms. *Int J Forecast*. 38(2):489–504.
6 Prasse B, Van Mieghem P. 2022. Predicting network dynamics without requiring the knowledge of the interaction graph. *Proc Natl Acad Sci USA*. 119(44):e2205517119.
7 Moss R, Zarebski AE, Carlson SJ, McCaw JM. 2019. Accounting for healthcare-seeking behaviours and testing practices in real-time influenza forecasts. *Trop Med Infect Dis*. 4(1):12.
8 Ibrahim NK. 2020. Epidemiologic surveillance for controlling COVID-19 pandemic: types, challenges and implications. *J Infect Public Health*. 13(11):1630–1638.
9 Clipman SJ, *et al*. 2022. Improvements in severe acute respiratory syndrome coronavirus 2 testing cascade in the united states: data from serial cross-sectional assessments. *Clin Infect Dis*. 74(9):1534–1542.
10 Rader B, *et al*. 2020. Geographic access to United States SARS-CoV-2 testing sites highlights healthcare disparities and may bias transmission estimates. *J Travel Med*. 27(7):taaa076.
11 Holden TM, *et al*. 2021. Geographic and demographic heterogeneity of SARS-CoV-2 diagnostic testing in Illinois, USA, March to December 2020. *BMC Public Health*. 21(1):1–13.
12 Runge M, *et al*. 2022. Modeling robust COVID-19 intensive care unit occupancy thresholds for imposing mitigation to prevent exceeding capacities. *PLOS Glob Public Health*. 2(5):e0000308.
13 Toh KB, Runge M, Richardson RAK, Hladish TJ, Gerardin J. 2023. Design of effective outpatient sentinel surveillance for COVID-19 decision-making: a modeling study. *BMC Infect Dis*. 23(1):287.
14 Pellis L, *et al*. 2021. Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. *Phil Trans R Soc B*. 376(1829):20200264.
15 Larremore DB, *et al*. 2021. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Sci Adv*. 7(1):eabd5393.
16 Marinović AB, Swaan C, van Steenbergen J, Kretzschmar M. 2015. Quantifying reporting timeliness to improve outbreak control. *Emerging Infect Dis*. 21(2):209–216.
17 Swaan C, van den Broek A, Kretzschmar M, Richardus JH. 2018. Timeliness of notification systems for infectious diseases: a systematic literature review. *PLoS One*. 13(6):e0198845.
18 Bastaki H, Carter J, Marston L, Cassell J, Rait G. 2018. Time delays in the diagnosis and treatment of malaria in non-endemic countries: a systematic review. *Travel Med Infect Dis*. 21:21–27.
19 Ahmed AE. 2017. Diagnostic delays in 537 symptomatic cases of Middle East respiratory syndrome coronavirus infection in Saudi Arabia. *Int J Infect Dis*. 62:47–51.
20 Sun K, Chen J, Viboud C. 2020. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *Lancet Digital Health*. 2(4):e201–e208.
21 Linton NM, *et al*. 2020. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *J Clin Med*. 9(2):538.

22 Lauer SA, *et al.* 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.
23 Kraemer MUG, *et al.* 2020. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*. 368(6490):493–497.
24 Leung K, Wu JT, Liu D, Leung GM. 2020. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet*. 395(10233):1382–1393.
25 Lin Q, *et al.* 2020. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *Int J Infect Dis*. 93:211–216.
26 Cereda D, *et al.* 2020. The early phase of the COVID-19 outbreak in Lombardy, Italy. *arXiv:2003.09320*.
27 Dehning J, *et al.* 2020. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*. 369(6500):eabb9789.
28 Günther F, Bender A, Katz K, Küchenhoff H, Höhle M. 2021. Nowcasting the COVID-19 pandemic in Bavaria. *Biom J*. 63(3): 490–502.
29 Tariq A, *et al.* 2020. Real-time monitoring the transmission potential of COVID-19 in Singapore, March 2020. *BMC Med*. 18(1):1–14.
30 Harris JE. 2022. Timely epidemic monitoring in the presence of reporting delays: anticipating the COVID-19 surge in New York city, September 2020. *BMC Public Health*. 22(1):871.
31 Génois M, Vestergaard CL, Cattuto C, Barrat A. 2015. Compensating for population sampling in simulations of epidemic spread on temporal contact networks. *Nat Commun*. 6(1):8860.
32 Mastrandrea R, Barrat A. 2016. How to estimate epidemic risk from incomplete contact diaries data? *PLoS Comput Biol*. 12(6): e1005002.
33 Sapienza A, Barrat A, Cattuto C, Gauvin L. 2018. Estimating the outcome of spreading processes on networks with incomplete information: a dimensionality reduction approach. *Phys Rev E*. 98(1):012317.
34 Lu FS, Nguyen AT, Link NB, Lipsitch M, Santillana M. 2021. Estimating the cumulative incidence of COVID-19 in the United States using influenza surveillance, virologic testing, and mortality data: Four complementary approaches. *PLOS Comput Biol*. 17(6):e1008994. https://doi.org/10.1371/journal.pcbi.1008994.
35 Aleta A, *et al.* 2020. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nat Hum Behav*. 4(9):964–971.
36 Cramer EY, *et al.* 2022. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the united states. *Proc Natl Acad Sci USA*. 119(15):e2113561119.
37 Chimmula VKR, Zhang L. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fract*. 135:109864.
38 Liu D, *et al.* 2020. A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. *arXiv:2004.04019*.
39 Rahimi I, Chen F, Gandomi AH. 2023. A review on COVID-19 forecasting models. *Neural Comput App*. 35(33):23671–23681. https://doi.org/10.1007/s00521-020-05626-8.
40 Chen J, *et al.* 2020. Clinical progression of patients with COVID-19 in Shanghai, China. *J Infect*. 80(5):e1–e6.
41 Voinsky I, Baristaite G, Gurwitz D. 2020. Effects of age and sex on recovery from COVID-19: analysis of 5769 Israeli patients. *J Infect*. 81(2):e102–e103.
42 Faes C, *et al.* 2020. Time between symptom onset, hospitalisation and recovery or death: statistical analysis of Belgian COVID-19 patients. *Int J Environ Res Public Health*. 17(20):7560.
43 Freeman GH. 1980. Fitting two-parameter discrete distributions to many data sets with one common parameter. *J R Stat Soc: Series C (Appl Stat)*. 29(3):259–267.
44 Johnson NL, Kemp AW, Kotz S. 2005. *Univariate discrete distributions*. New York (NY): John Wiley & Sons444.
45 Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. *J Mach Learn Res*. 13(2):281–305.
46 Sun K, *et al.* 2021. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science*. 371(6526):eabe2424.
47 Kretzschmar ME, *et al.* 2020. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *Lancet Public Health*. 5(8):e452–e459.
48 https://www.agi.it/cronaca/news/2020-04-28/fase-2-arcuri-tamponi-mascherine-app-8461502/, 2020.
49 Jespers V, *et al.* 2020. International comparison of COVID-19 testing and contact tracing strategies. Fps Health, Food Chain Safety and Environment: Lieven De Raedt Sciensano: Ana Hoxha. COVID-19 KCE contributions. 29.
50 https://m.chinanews.com/wap/detail/zw/sh/2020/01-28/9071697.shtml, 2020.
51 Fu H, *et al.* 2021. Database of epidemic trends and control measures during the first wave of COVID-19 in mainland China. *Int J Infect Dis*. 102:463–471.
52 https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/2020-04-08-en.pdf?__blob=publicationFile, 2020.
53 Akbarov A, Wu S. 2013. Warranty claims data analysis considering sales delay. *Qual Reliab Eng Int*. 29(1):113–123.
54 Nesteruk I. 2021. *COVID-19 pandemic dynamics: mathematical simulations*. Singapore: Springer Nature.
55 Ord J. K. 1974. Families of frequency distributions. *Int Stat Rev*. 42(2):231. https://doi.org/10.2307/1403084.

Supplementary Information for

Reporting delays: a widely neglected impact factor in COVID-19 forecasts

Long Ma, Zhihao Qiu, Piet Van Mieghem, Maksim Kitsak

Maksim Kitsak
E-mail: M.A.Kitsak@tudelft.nl

This PDF file includes:

Supporting Information Text

A. Data availability. The data sources of COVID-19 cases for each country/region are as follows:
1. **Italy**: The data was collected daily from Dipartimento della Protezione Civile (http://www.salute.gov.it/portale/home.html) from Feb 21, 2020 to May 4, 2020. The available data can be found in WHO COVID-19 global dataset https://covid19.who.int/WHO-COVID-19-global-data.csv, https://opendatamds.maps.arcgis.com/apps/dashboards/12a643195a994a558f4dbd603338ee33, https://www.worldometers.info/coronavirus/ and https://it.wikipedia.org/wiki/Statistiche_della_pandemia_di_COVID-19_in_Italia
2. **Spain**: The data was collected daily from Ministry of Health (Spain) (https://www.sanidad.gob.es/home.htm) from February 25, 2020 to May 15, 2020. The available data can be found in WHO COVID-19 global dataset https://covid19.who.int/WHO-COVID-19-global-data.csv , https://www.rtve.es/noticias/20230331/mapa-del-coronavirus-espana/2004681.shtml, https://www.worldometers.info/coronavirus/ and https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Spain#Statistics
3. **Wuhan**: The data was collected daily from National Health Commission of the People's Republic of China (http://en.nhc.gov.cn/) from January 18, 2020 to March 15, 2020. The available data can be found in Wuhan Municipal Health Commission https://wjw.wuhan.gov.cn/ztzl_28/fk/yqtb/
4. **Turkey**: The data was collected daily from Ministry of Health (Turkey) (https://www.saglik.gov.tr/) from March 11, 2020 to May 16, 2020. The available data can be found in WHO COVID-19 global dataset https://covid19.who.int/WHO-COVID-19-global-data.csv, https://www.worldometers.info/coronavirus/ and https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Turkey#Statistics
5. **Hubei**: The data was collected daily from National Health Commission of the People's Republic of China (http://en.nhc.gov.cn/) from January 18, 2020 and end on March 15, 2020. The available data can be found in http://www.nhc.gov.cn/yjb/s7860/202205/4b6b9808ba8a469c9026f05d3f35546e.shtml and https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data/China_medical_cases_by_province
6. **Romania**: The data was collected daily from Ministry of Health (Romania) (http://www.ms.ro/comunicate/) from February 26, 2020 to May 23, 2020. The available data can be found in WHO COVID-19 global dataset https://covid19.who.int/WHO-COVID-19-global-data.csv, https://www.worldometers.info/coronavirus/ and https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Romania#Cumulative_cases
7. **Germany**: The data was collected daily from Robert Koch-Institut (RKI) (https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Fallzahlen.html) and new situation reports of the RKI (https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html). from March 1, 2020 to May 19, 2020. The available data can be found in WHO COVID-19 global dataset https://covid19.who.int/WHO-COVID-19-global-data.csv, https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Nov_2020/2020-11-11-de.pdf?__blob=publicationFile, https://www.worldometers.info/coronavirus/ and https://en.wikipedia.org/wiki/Statistics_of_the_COVID-19_pandemic_in_Germany.
8. **Denmark**: The data was collected daily from Statens Serum Institut (https://en.ssi.dk/) from February 27, 2020 to May 16, 2020. The available data can be found in WHO COVID-19 global dataset https://covid19.who.int/WHO-COVID-19-global-data.csv, https://covid19.datelazi.ro/, https://www.worldometers.info/coronavirus/ and https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Denmark#Statistics.

We extract the epidemic data time series from these datasets and make them available at https://doi.org/10.6084/m9.figshare.22639519.v1.

B. Qualitative explanation for the formation mechanism of the loop patterns in epidemic data. In the main text, we consider the epidemic data reported for Spain during the first wave of the COVID-19 pandemic, Fig. 1. We attribute the observed misalignment of peaks of the epidemic data as well as the loop patterns to the presence of delays in the reporting data. Our observations are not specific to Spain. As evidenced in Fig. S1 and Fig. S2, we observe similar patterns in all 8 regions of interest.

To verify if the observed patterns could follow from reporting delays, we conduct a test using synthetic epidemic data. In more precise terms, we generated synthetic epidemic time series using the SIRD epidemic model, Eq. [3], setting epidemic parameters to $\beta = 0.5$, $\gamma_r = 0.2$, and $\gamma_d = 0.05$. After obtaining synthetic epidemic times series $Y[k]$, we added reporting delays using Eq. [1]. We assumed that reporting delays follow Pólya-Aeppli distributions with $\zeta_D = 1/3$, $\mu_D = 2$, $\zeta_I = 2$, $\mu_I = 2$, $\zeta_R = 3$, and $\mu_R = 6$, where sub-indices D, I, and R refer to deceased, infected, and recovered sub-populations. These parameters correspond to distributions with $E[T_D] = 2/3$, $\mathrm{Var}[T_D] = 10/9$, $E[T_I] = 4$, $\mathrm{Var}[T_I] = 20$, $E[T_R] = 18$, and $\mathrm{Var}[T_R] = 342$, see Fig. S3(a). Fig. S3(b,c) illustrates that the resulting synthetic data exhibit patterns very similar to those observed for the COVID-19 data.

The formation of the $\Delta\widetilde{R}[k+1]$ versus $\tilde{I}[k]$ and $\Delta\widetilde{D}[k+1]$ versus $\widetilde{I}[k]$ loop patterns in Fig. S3(c) is due to the effective horizontal shifts of the corresponding times series due to reporting delays. Indeed, let us split the observation time window into three windows formed by the maxima of the $\Delta\widetilde{D}[k+1]$ and $\widetilde{I}[k]$ curves, as shown in Fig. S3(b). In window i, both $\widetilde{I}[k]$ and $\Delta\widetilde{D}[k+1]$ increase as a function of discrete time step k. Since reporting delays in the synthetic $\Delta\widetilde{D}[k+1]$ data are smaller than those in $\widetilde{I}[k]$, this time window corresponds to the upper branch of the $\Delta\widetilde{D}[k+1]$ versus $\widetilde{I}[k]$ loop in Fig. S3(c). In window ii, $\widetilde{I}[k]$ increases while $\Delta\widetilde{D}[k+1]$ decreases. Thus, window ii corresponds to the top (decreasing) section of the $\Delta\widetilde{D}[k+1]$ versus $\widetilde{I}[k]$ loop, Fig. S3(c). Finally, in window iii both $\Delta\widetilde{D}[k+1]$ and $\widetilde{I}[k]$ decrease as a function of time step k resulting in the lowest section of the $\Delta\widetilde{D}[k+1]$ versus $\widetilde{I}[k]$ loop, Fig. S3(c). Combined, all sections correspond to the loop pattern of Fig. S3(c). with points progressing in the clockwise direction. Similar considerations explain the counterclockwise loop pattern in the $\Delta\widetilde{R}[k+1]$ vs $\tilde{I}[k]$ scatter plot. The counterclockwise progression of points in this loop pattern is due to $\Delta\widetilde{D}[k+1]$ lagging behind the $\tilde{I}[k]$ time series.

Regions	λ_I	θ_I	λ_R	θ_R	λ_D	θ_D
Italy	2.8741	0.4193	2.0701	0.0719	0.1787	0.6750
Spain	1.0803	0.1632	1.2673	0.0553	0.1713	0.6574
Wuhan	0.4661	0.0186	2.6485	0.1377	0.0681	0.0154
Turkey	0.6803	0.0264	2.6537	0.1763	0.0135	0.0146
Hubei	0.1745	0.0075	2.2779	0.1866	0.0033	0.0015
Romania	0.3902	0.0204	0.0026	0.0008	0.0632	0.8985
Germany	0.4122	0.0230	0.0028	0.0007	0.1878	0.1371
Denmark	1.0399	0.0387	0.8382	0.4240	0.0081	0.0045

Table S1. Inferred parameters of the reporting delays under the assumption of the Pólya-Aeppli distribution.

Country/Region	Before, $O_b(\widetilde{Y})$	After, $O_b(Y)$	Relative change, $(O_b(Y) - O_b(\widetilde{Y}))/O_b(\widetilde{Y})$
Italy	0.24	0.84	2.5
Spain	0.36	0.98	1.7
Wuhan	0.06	0.63	9.5
Turkey	0.26	0.85	2.5
Hubei	0.10	0.71	6.1
Romania	0.36	0.70	0.94
Germany	0.37	0.76	1.1
Denmark	0.31	0.72	1.3

Table S2. Relative changes in the objective function $O_b(\widetilde{Y})$ before and after the removal of reporting delays.

C. Forecast the COVID-19 pandemic. The COVID-19 pandemic is predicted based on the Algorithms 1 and 2. Algorithm 1 is to forecast the future infected fractions without considering the effect of reporting delays. Algorithm 2 is to forecast the future infected fractions considering the reporting delays.

Algorithm 1 `Epidemic forecast without removal of reporting delays, experiment 1.`

Input: fraction of daily reported infected cases $\Delta\tilde{I}[1], ..., \Delta\tilde{I}[n]$.
Output: predicted fraction of infections $\Delta\hat{I}[n+1], ..., \Delta\hat{I}[n+n_{\text{pred}}]$.
Smooth the data by Matlab toolbox *smoothdata*.
Set the initial value of the loss function $\Theta_s \leftarrow 1$; the initial infection rate $\beta_s \leftarrow 0.01$; the initial removed rate (the sum of recovered rate and deceased rate) $\gamma_s \leftarrow 0.01$; the initial infected fraction $I_s[0] \leftarrow 0.01$;
for $\beta = 0.01, 0.02, ..., 1$ **do**
    for $\gamma = 0.01, 0.02, ..., 1$ **do**
        for $k = 0, 1, ..., 100$ **do**
            $I[0] = 10^{-2-k/25}$;
            Numerically solve the equations of the SIRD model based on the parameters $I[0]$, β and γ and obtain an infection curve $\Delta\mathcal{I}[1], ..., \Delta\mathcal{I}[n+n_p]$.
            Scale the simulated curve $\Delta\mathcal{I}[1], ..., \Delta\mathcal{I}[n+n_p]$ to data $\Delta\hat{I}[1], ..., \Delta\hat{I}[n+n_p]$ by letting $\Delta\hat{I}[i] = \Delta\mathcal{I}[i] \times \Delta\tilde{I}[1]/\Delta\mathcal{I}[1]$ for day $i - 1, 2, \cdots, n+n_p$.
            Calculate the lose function $\Theta = \sqrt{\frac{1}{n}\sum_{k=0}^{n-1}(\Delta\hat{I}[k] - \Delta\tilde{I}[k])^2}$.
            if $\Theta < \Theta_s$ **then**
                $\Theta_s \leftarrow \Theta$; $I_s[0] \leftarrow I[0]$; $\beta_s \leftarrow \beta$; $\gamma_s \leftarrow \gamma$.
Obtain the best forecast results based on the optimized parameters $I_s[0]$, β_s and γ_s.

Algorithm 2 `Forecast the real pandemic considering the reporting delays`

Input: fraction of daily reported infected cases $\Delta\tilde{I}[1], ..., \Delta\tilde{I}[n]$; fraction of daily reported deceased cases $\Delta\tilde{D}[1], ..., \Delta\tilde{D}[n]$; fraction of daily reported recovered cases $\Delta\tilde{R}[1], ..., \Delta\tilde{R}[n]$; prediction time n_{pred}.
Output: predicted fraction of infections $\Delta\hat{I}[n+1], ..., \Delta\hat{I}[n+n_{\text{pred}}]$.
Smooth the data by Matlab toolbox *smoothdata*.
Infer delay distribution for infected cases T_I using the reported data $\Delta\tilde{I}$, $\Delta\tilde{R}$ and $\Delta\tilde{D}$.
Obtain the inferred data $\Delta\bar{I}[1], ..., \Delta\bar{I}[n]$ by removing the effect of reporting delays.
Set the initial value of the loss function $\Theta_s \leftarrow 1$; the initial infection rate $\beta_s \leftarrow 0.01$; the initial removed rate (the sum of recovered rate and deceased rate) $\gamma_s \leftarrow 0.01$; the initial infected fraction $I_s[0] \leftarrow 0.01$;
for $\beta = 0.01, 0.02, ..., 1$ **do**
    for $\gamma = 0.01, 0.02, ..., 1$ **do**
        for $k = 0, 1, ..., 100$ **do**
            $I[0] = 10^{-2-k/25}$;
            Numerically solve the equations of SIR model based on the parameters $I[0]$, β and γ and obtain an infection curve $\Delta\mathcal{I}[1], ..., \Delta\mathcal{I}[n+n_p]$.
            Scale the simulated curve $\Delta\mathcal{I}[1], ..., \Delta\mathcal{I}[n+n_p]$ to data $\Delta\hat{I}[1], ..., \Delta\hat{I}[n+n_p]$ by letting $\Delta\hat{I}[i] = \Delta\mathcal{I}[i] \times \Delta\bar{I}[1]/\Delta\mathcal{I}[1]$ for day $i - 1, 2, \cdots, n+n_p$.
            Calculate the lose function $\Theta = \sqrt{\frac{1}{n}\sum_{k=0}^{n-1}(\Delta\hat{I}[k] - \Delta\bar{I}[k])^2}$.
            if $\Theta < \Theta_s$ **then**
                $\Theta_s \leftarrow \Theta$; $I_s[0] \leftarrow I[0]$; $\beta_s \leftarrow \beta$; $\gamma_s \leftarrow \gamma$.
Obtain the best forecast results based on the optimized parameters $I_s[0]$, β_s and γ_s.
Add the reporting delays to the forecast data using Eq. [1]

D. Governing equation of the time shift. The fraction $Y[k]$ is contained in $[0, 1]$ and measured per day. Hence, $Y[k]$ is a real-valued random variable and the analysis is in discrete time k, where k is an integer that represents day $k - k_0$ since the start of the measurements of the fraction Y at day $k_0 \in \mathbb{Z}$. In the sequel, as in (1), capital letter refer to random variable and small letters to numbers.

Our basic observation lies in a time delay T between the *real* value $Y[k]$ and the reported value $Y_{\text{rep}}[k]$, which translates to

$$Y[k] = Y_{\text{rep}}[k+T] \quad [1]$$

Here, we implicitly assume that the reported value $Y_{\text{rep}}[k]$ exactly measures the real fraction and only differs from reality in the time delay T. If errors or uncertainties appear, then we can replace $Y_{\text{rep}}[k]$ by $Y_{\text{rep}}[k] = \widetilde{Y}_{\text{rep}}[k] + U[k]$, where U reflects

the uncertainty or error in reporting or measuring. The delay T must be non-negative, i.e. $T \geq 0$, because the real event $Y[k]$ occurs at discrete time k and its reporting occurs at $k+T \geq k$, which cannot be earlier than the time k. Moreover, the delay T is also an integer because we assume a discrete-time analysis, else Eq. [1] demands us to take the integer value $[T]$ of $T = [T] + \langle T \rangle$, where the fractional part $0 \leq \langle T \rangle < 1$. We avoid that complication and consider T as a discrete random variable. If T is discrete, then all involved random variables are discrete. Before proceeding, we thus approximate $Y(t)$ at continuous time $k-1 < t \leq k$, by $Y[k] = \int_{k-1}^{k} Y(u)\, du$. Hence, the instantaneous fraction $Y(t)$, denoted by "round" bracket (.), and cumulative fraction $Y[k]$ in one timeslot, denoted by "square" brackets [.], are physically different random variables! Furthermore, the mean-field approximation (as in Eq. [1] of the main text) only writes the equations for the *average* fraction of infected nodes and we refer to (1, Sec. 17.3.2; Sec. 17.4) for the relation between the Markov process and its mean-field approximation.

The basic observation in [1] contains two intertwined random variables, namely the reported fraction Y_{rep} and the delay T. Hence, we use the law of total probability (1, p. 23) and conditioning,

$$\Pr[Y_{\text{rep}}[k+T] \leq y] = \sum_{m=0}^{\infty} \Pr[Y_{\text{rep}}[k+T] \leq y | T = m] \Pr[T = m]$$

Since we confine to a mean-field analysis and are only interested in the *average* fraction of infected (see eq. (1,2) in the main text), we better take the expectation operator $E[.]$ instead of the probability Pr[.] operator:

$$E[Y_{\text{rep}}[k+T]] = \sum_{m=0}^{\infty} E[Y_{\text{rep}}[k+T] | T = m] \Pr[T = m] \quad [2]$$

which is readily obtained from the former equation by using the definition of the mean (e.g. (1, (2.36) on p. 18)), namely $E[Y_{\text{rep}}[k+T]] = \int_0^1 \Pr[Y_{\text{rep}}[k+T] > y]\, dy$.

This last equation [2] involves the conditional expectations $E[Y_{\text{rep}}[k+T] | T = m]$. If we assume that T is independent of Y_{rep}, then $E[Y_{\text{rep}}[k+T] | T = m] = E[Y_{\text{rep}}[k+m]]$. After taking the expectation of the hypothesis [1] and combining with [2], assuming independence between fraction Y and delay T, then we obtain

$$E[Y[k]] = \sum_{m=0}^{\infty} E[Y_{\text{rep}}[k+m]] \Pr[T = m]$$

Finally, if Y reflects the fraction of infected (and similarly for the other compartments in epidemic models), then we simplify the notation as $I[k] = E[Y[k]]$ and arrive at our approximative observation hypothesis for the average fraction of infected individuals

$$I[k] = \sum_{m=0}^{\infty} I_{\text{rep}}[k+m] \Pr[T = m] \quad [3]$$

In summary, if the time delay T and the reported value Y_{rep} are independent, then the observation hypothesis [1] translates the average fraction of infected (similarly for removed and deceased) to the weighted sum of the average fraction of the reported, where the weight is the probability $\Pr[T = m]$ that the (integer) delay T equals m. If independence does not hold, the above derivation shows that it is difficult to compute the relation between $Y_{\text{rep}}[k]$ and $Y[k]$, unless the conditional probabilities in [2] can be determined. We argue that independence between T and Y_{rep} is reasonable, although both random variables are weakly positively correlated. Indeed, the larger the fraction Y, the more people need to be checked and the longer the reporting may take. If the checking capacity is sufficiently large, we may assume approximate independence.

D.1. SIR compartments. The discrete-time SIR model is defined (see e.g. (2)), for every node i in the graph G at discrete time k, by viral state vector $v_i[k] = (S_i[k], I_i[k], R_i[k])$. The governing SIR equations in discrete time as

$$I_i[k+1] = (1-\delta_i) I_i[k] + (1 - I_i[k] - R_i[k]) \sum_{n=1}^{N} \beta_{in} I_n[k] \quad [4]$$

$$R_i[k+1] = R_i[k] + \delta_i I_i[k] \quad [5]$$

and the fraction of susceptible individuals follows from the conservation law as

$$S_i[k] = 1 - I_i[k] - R_i[k]$$

Here, β_{in} denotes the infection probability from node i to node n and δ_i is the curing probability of node i. With the initial condition $R_i[0] = 0$, the linear difference equation [5] is readily solved for $k > 0$

$$R_i[k] = \delta_i \sum_{m=0}^{k-1} I_i[m]$$

and [4] reduces to

$$I_i[k+1] - (1-\delta_i) I_i[k] = (1 - I_i[k] - \delta_i \sum_{m=0}^{k-1} I_i[m]) \left(\beta_{ii} I_i[k] + \sum_{n=1; n \neq i}^{N} \beta_{in} I_n[k] \right) \quad [6]$$

D.2. The time delays T_S, T_I and T_S are correlated. Relation [2] holds for any fraction Y. If we confine to the SIR model with three compartments, then the approximate equations [3] are

$$\left\{\begin{array}{c} S[k]=\sum_{m=0}^{\infty} S_{\mathrm{rep}}[k+m]\Pr[T_S=m] \\ I[k]=\sum_{m=0}^{\infty} I_{\mathrm{rep}}[k+m]\Pr[T_I=m] \\ R[k]=\sum_{m=0}^{\infty} R_{\mathrm{rep}}[k+m]\Pr[T_R=m] \end{array}\right.$$

The compartmental conservation law $S[k]+I[k]+R[k]=1$ (and also $S_{\mathrm{rep}}[k]+I_{\mathrm{rep}}[k]+R_{\mathrm{rep}}[k]=1$) leads to

$$\begin{aligned} 1 &= \sum_{m=0}^{\infty}\left(1-I_{\mathrm{rep}}[k+m]-R_{\mathrm{rep}}[k+m]\right)\Pr[T_S=m]+\sum_{m=0}^{\infty} I_{\mathrm{rep}}[k+m]\Pr[T_I=m] \\ &\quad +\sum_{m=0}^{\infty} R_{\mathrm{rep}}[k+m]\Pr[T_R=m] \end{aligned}$$

or

$$\sum_{m=0}^{\infty} I_{\mathrm{rep}}[k+m]\left(\Pr[T_S=m]-\Pr[T_I=m]\right)=\sum_{m=0}^{\infty} R_{\mathrm{rep}}[k+m]\left(\Pr[T_R=m]-\Pr[T_S=m]\right)$$

This relation illustrates that the timeshifts T_S, T_I and T_R are correlated. The correlation or dependence is expected because the timeshifts T_S, T_I and T_R are all obtained from a same measurement or reporting procedure. Interestingly, if we assume that T_S, T_I and T_R have the same distribution, $\Pr[T_S=m]=\Pr[T_I=m]=\Pr[T_R=m]$, (which does not imply that the timeshifts, i.e. the random variables, are the same!, i.e. $T_S \neq T_I \neq T_R$), then the compartmental conservation is always satisfied.

D.3. Set of quadratic equations for $\Pr[T_I=m]$. Introducing the observation hypothesis $I_i[k]=\sum_{m=0}^{\infty} I_{\mathrm{rep,i}}[k+m]\Pr[T_{I,i}=m]$ in [3] at node i into [6] yields

$$\begin{aligned} Q_L[k] &= \sum_{m=0}^{\infty}\left(I_{\mathrm{rep,i}}[k+1+m]-(1-\delta_i)I_{\mathrm{rep,i}}[k+m]\right)\Pr[T_{I,i}=m] \\ Q_R[k] &= \left(1-\sum_{m=0}^{\infty}\left\{I_{\mathrm{rep,i}}[k+m]-\delta_i\sum_{l=0}^{k-1} I_{\mathrm{rep,i}}[l+m]\right\}\Pr[T_{I,i}=m]\right) \\ &\quad \left(\beta_{ii}\sum_{m=0}^{\infty} I_{\mathrm{rep,i}}[k+m]\Pr[T_{I,i}=m]+\sum_{n=1;n\neq i}^{N}\beta_{in}\sum_{m=0}^{\infty} I_{\mathrm{rep,n}}[k+m]\Pr[T_{I,n}=m]\right) \end{aligned}$$

where the left-hand side Q_L and right-hand side Q_R of equation [6] are equal, i.e. $Q_L[k]=Q_R[k]$. We omit in the sequel the index R and L. We can proceed if we assume that the time-delay in different nodes has the same distribution, i.e. $\Pr[T_{I,i}=m]=\Pr[T_I=m]$ for each node i in the contact graph. Then,

$$\begin{aligned} Q[k] &= \sum_{m=0}^{\infty}\left(I_{\mathrm{rep,i}}[k+1+m]-(1-\delta_i)I_{\mathrm{rep,i}}[k+m]\right)\Pr[T_I=m] \\ &= \sum_{m=0}^{\infty}\left\{1-I_{\mathrm{rep,i}}[k+m]-\delta_i\sum_{l=0}^{k-1} I_{\mathrm{rep,i}}[l+m]\right\}\Pr[T_I=m]\sum_{r=0}^{\infty}\left(\sum_{n=1}^{N}\beta_{in}I_{\mathrm{rep,n}}[k+r]\right)\Pr[T_I=r] \end{aligned}$$

We rewrite the double sum

$$\begin{aligned} \sum_{m=0}^{\infty}\sum_{r=0}^{\infty} f(m)\,g(r) &= \sum_{m=0}^{\infty}\sum_{r=0}^{\infty}\left\{1-I_{\mathrm{rep,i}}[k+m]-\delta_i\sum_{l=0}^{k-1} I_{\mathrm{rep,i}}[l+m]\right\}\left(\sum_{n=1}^{N}\beta_{in}I_{\mathrm{rep,n}}[k+r]\right) \\ &\quad \times\Pr[T_I=m]\Pr[T_I=r] \end{aligned}$$

as

$$\sum_{m=0}^{\infty}\sum_{r=0}^{\infty} f(m)\,g(r)=\sum_{m=0}^{\infty} f(m)\sum_{r=0}^{m-1} g(r)+\sum_{m=0}^{\infty} f(m)\sum_{r=m}^{\infty} g(r)$$

Reversing the summations in the last sum,

$$\sum_{m=0}^{\infty} f(m)\sum_{r=m}^{\infty} g(r)=\sum_{r=0}^{\infty} g(r)\sum_{m=0}^{r} f(m)=\sum_{r=0}^{\infty} g(r)\,f(r)+\sum_{r=0}^{\infty} g(r)\sum_{m=0}^{r-1} f(m)$$

Hence, we obtain (after interchanging the indices in the last)

$$\sum_{m=0}^{\infty}\sum_{r=0}^{\infty} f(m)\, g(r) = \sum_{r=0}^{\infty} g(r)\, f(r) + \sum_{m=0}^{\infty}\sum_{r=0}^{m-1} \{f(m)\, g(r) + g(m)\, f(r)\}$$

Finally, for each discrete time $0 \leq k$, we obtain

$$\begin{aligned} Q[k] &= \sum_{m=0}^{\infty} \left(I_{\text{rep,i}}[k+1+m] - (1-\delta_i) I_{\text{rep,i}}[k+m]\right) \Pr[T_I = m] \\ &= \sum_{m=0}^{\infty} \left(\sum_{n=1}^{N} \beta_{in} I_{\text{rep,n}}[k+m]\right) \left(1 - I_{\text{rep,i}}[k+m] - \delta_i \sum_{l=0}^{k-1} I_{\text{rep,i}}[l+m]\right) \left(\Pr[T_I = m]\right)^2 \\ &\quad + \sum_{m=0}^{\infty} \Pr[T_I = m] \sum_{r=0}^{m-1} \Pr[T_I = r] \left\{ \begin{array}{c} \sum_{n=1}^{N} \beta_{in} I_{\text{rep,n}}[k+r] \left(1 - I_{\text{rep,i}}[k+m] - \delta_i \sum_{l=0}^{k-1} I_{\text{rep,i}}[l+m]\right) \\ + \sum_{n=1}^{N} \beta_{in} I_{\text{rep,n}}[k+m] \left(1 - I_{\text{rep,i}}[k+r] - \delta_i \sum_{l=0}^{k-1} I_{\text{rep,i}}[l+r]\right) \end{array} \right\} \end{aligned}$$

In order to limit the infinite sum, we can choose a time point M and assume that $\Pr[T_I = m] = 0$ for $m > M$. Denoting $x[m] = \Pr[T_I = m]$, the above equation possesses the form

$$\sum_{m=0}^{M} A_m[k]\, x[m] = \sum_{m=0}^{M} B_m[k]\, C_m[k]\, (x[m])^2 + \sum_{m=0}^{M} x[m] \sum_{r=0}^{m-1} x[r] \{B_r[k]\, C_m[k] + B_m[k]\, C_r[k]\}$$

which leads to a set of quadratic equations in the variables $\{x[m]\}_{0 \leq m \leq M}$ for each $k \geq 0$, that can be solved numerically, together with $\sum_{m=0}^{M} x[m] = 1$.

In summary, assuming an SIR epidemics with the knowledge of the infection probabilities β_{in} between node i and n and the nodal curing probability δ_i at each node i, then the set $\{I_{\text{rep,n}}[k]\}_{k \geq k_0}$ of reported fractions of infected $I_{\text{rep,n}}[k]$ at discrete time k and at each node n is sufficient to compute the delay distribution $\Pr[T_I = m]$ (where we have assumed that the distribution $\Pr[T_{I,n} = m] = \Pr[T_I = m]$ is the same for each node). The resulting set of quadratic equation is rather demanding to solve and justifies the approach in the main text: we assume that, within a broad class of probability distributions with two parameters, a certain range of those parameters satisfies the quadratic set for $\Pr[T_I = m]$.

References

1. P. Van Mieghem. *Performance Analysis of Complex Networks and Systems.* Cambridge University Press, Cambridge, U.K., 2014.
2. B. Prasse, M. A. Achterberg, L. Ma, and P. Van Mieghem. Network inference-based prediction of the COVID-19 outbreak in the Chinese province Hubei. *Applied Network Science*, to appear, also on arXiv:2002.04482 2020.
3. P. Van Mieghem. The asymptotic behaviour of queueing systems: Large deviations theory and dominant pole approximation. *Queueing Systems*, 23:27–55, 1996.
4. E. C. Titchmarsh. *The Theory of Functions.* Oxford University Press, Amen House, London, 1964.

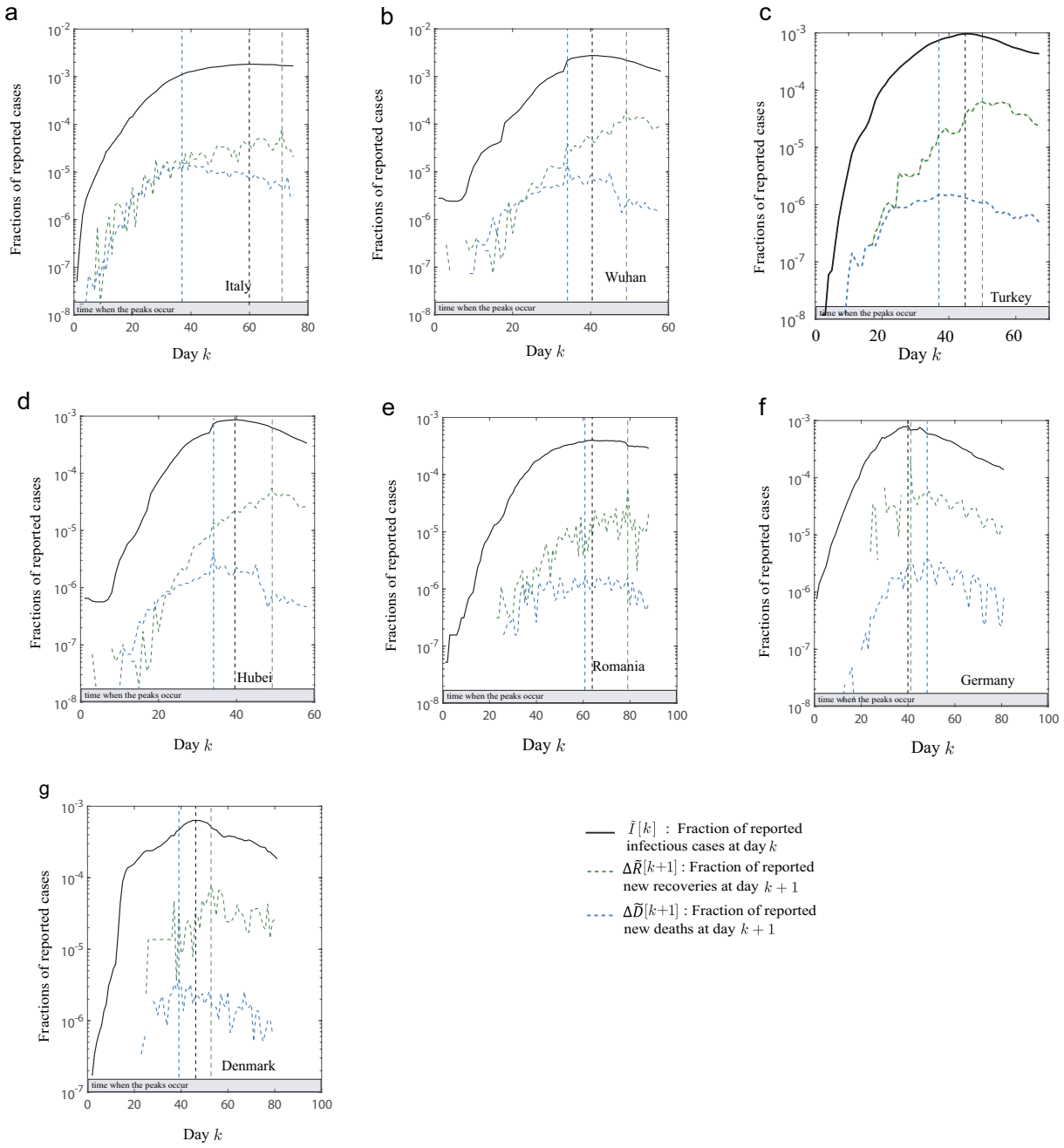

Fig. S1. Time series of fractions of reported active infections $\tilde{I}[k]$, new reported recoveries $\Delta\widetilde{R}[k+1]$ and new reported deaths $\Delta\widetilde{D}[k+1]$ for (a) Italy, (b) Wuhan province, (c) Turkey, (d) Hubei province, (e) Romania, (f) Germany, and (g) Denmark. Vertical dashed lines highlight the locations of data peaks.

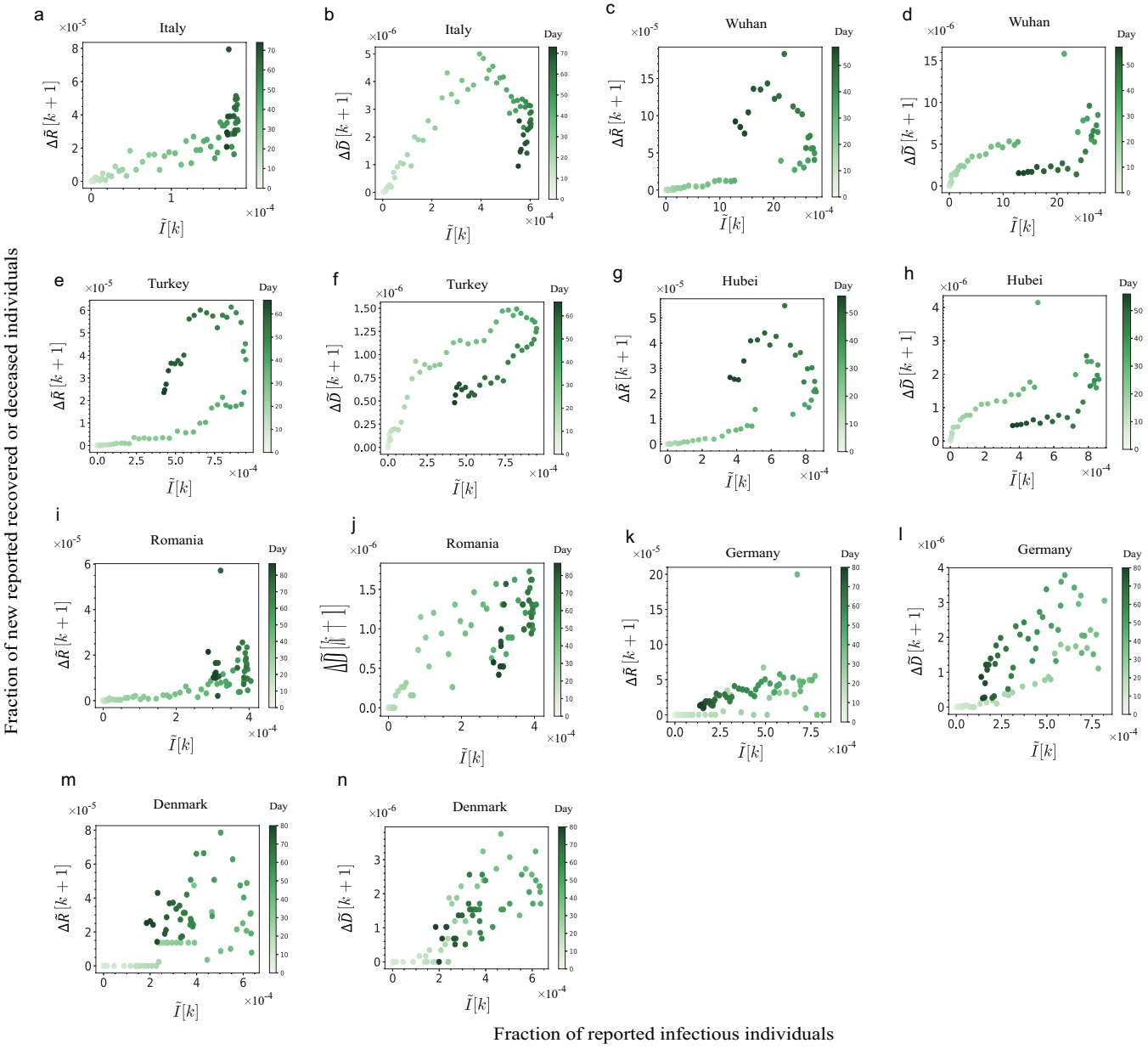

Fig. S2. Panels display pairwise color-coded scatter plots of the fraction of reported active infections $\widetilde{I}[k]$ as a function of the fraction of new reported recoveries $\Delta\widetilde{R}[k+1]$ and the fraction of reported active infections $\widetilde{I}[k]$ as a function of the fraction of reported deaths $\Delta\widetilde{D}[k+1]$ for (a,b) Italy, (c,d) Wuhan province, (e,f) Turkey, (g,h) Hubei province, (i,j) Romania, (k,l) Germany, and (m,n) Denmark. Colors, from light to dark green, reflect different days in the data ranging, respectively, from $k=0$ to $k=57$. Note that data points for $\widetilde{I}[k]$ versus $\Delta\widetilde{R}[k+1]$ evolve in the counter-clockwise direction, while data points for $\widetilde{I}[k]$ versus $\Delta\widetilde{D}[k+1]$ evolve in the clockwise direction.

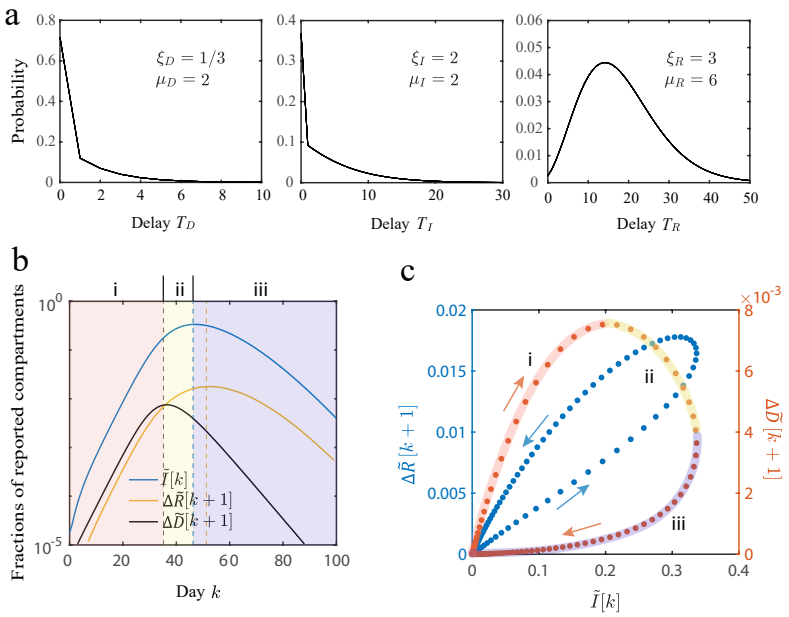

Fig. S3. Effects of reporting delays on synthetic data. (a) Pólya-Aeppli distributions generating reporting delays in the deceased, infected and recovered datasets. (b) Epidemic data generated with the SIRD model containing reporting delays. (c) Changes in the fraction of recovered $\Delta R[k+1]$ and deceased $\Delta D[k+1]$ individuals as a function of the fraction of infected individuals $I[k]$. Note that reporting delays lead to the appearance of the clockwise and counterclockwise loop patterns in the epidemic data.

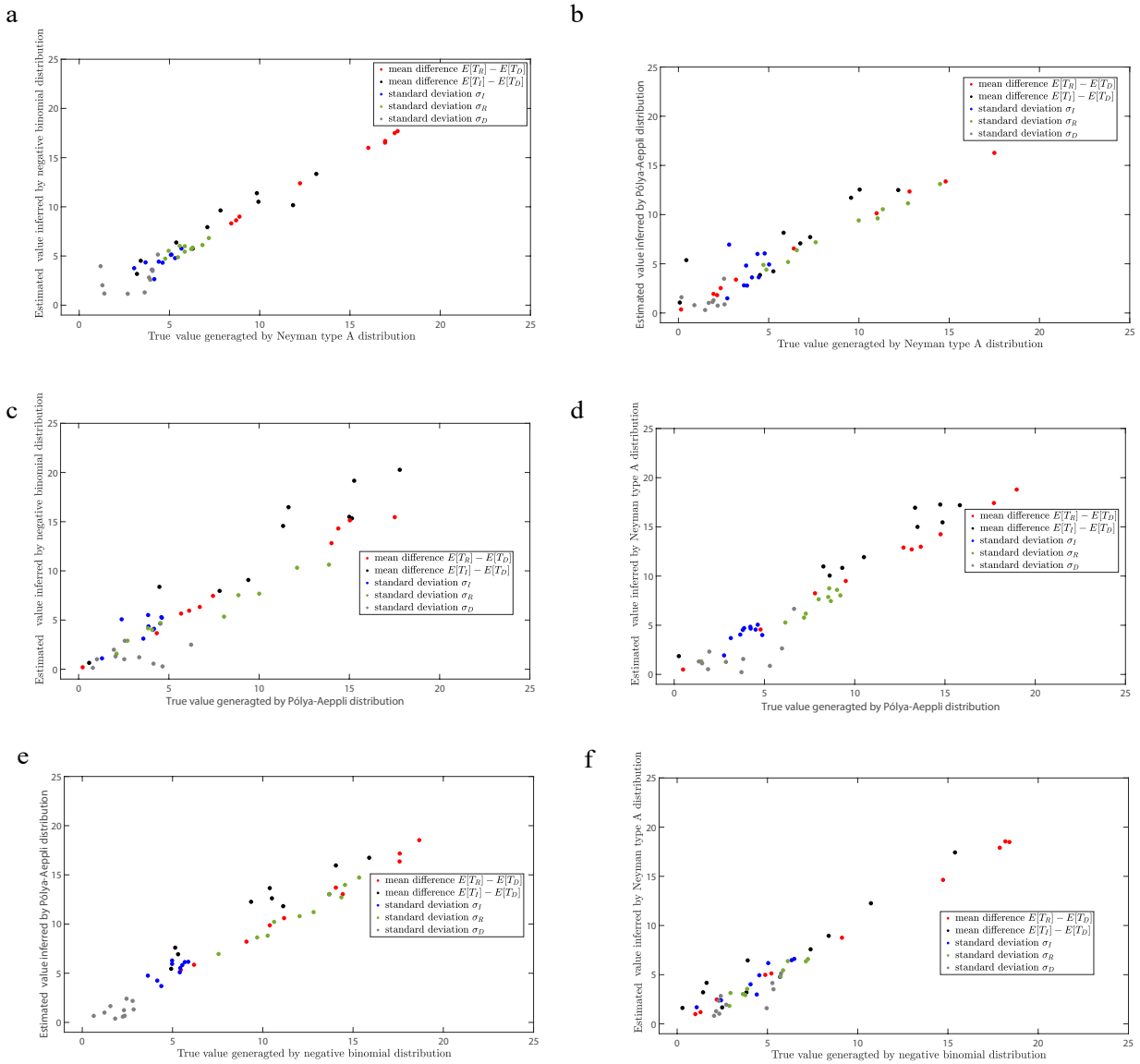

Fig. S4. Inference of reporting delays on synthetic data. This figure will feature 6 panels corresponding to all pairs of distributions: data generated with X, inferred with Y.

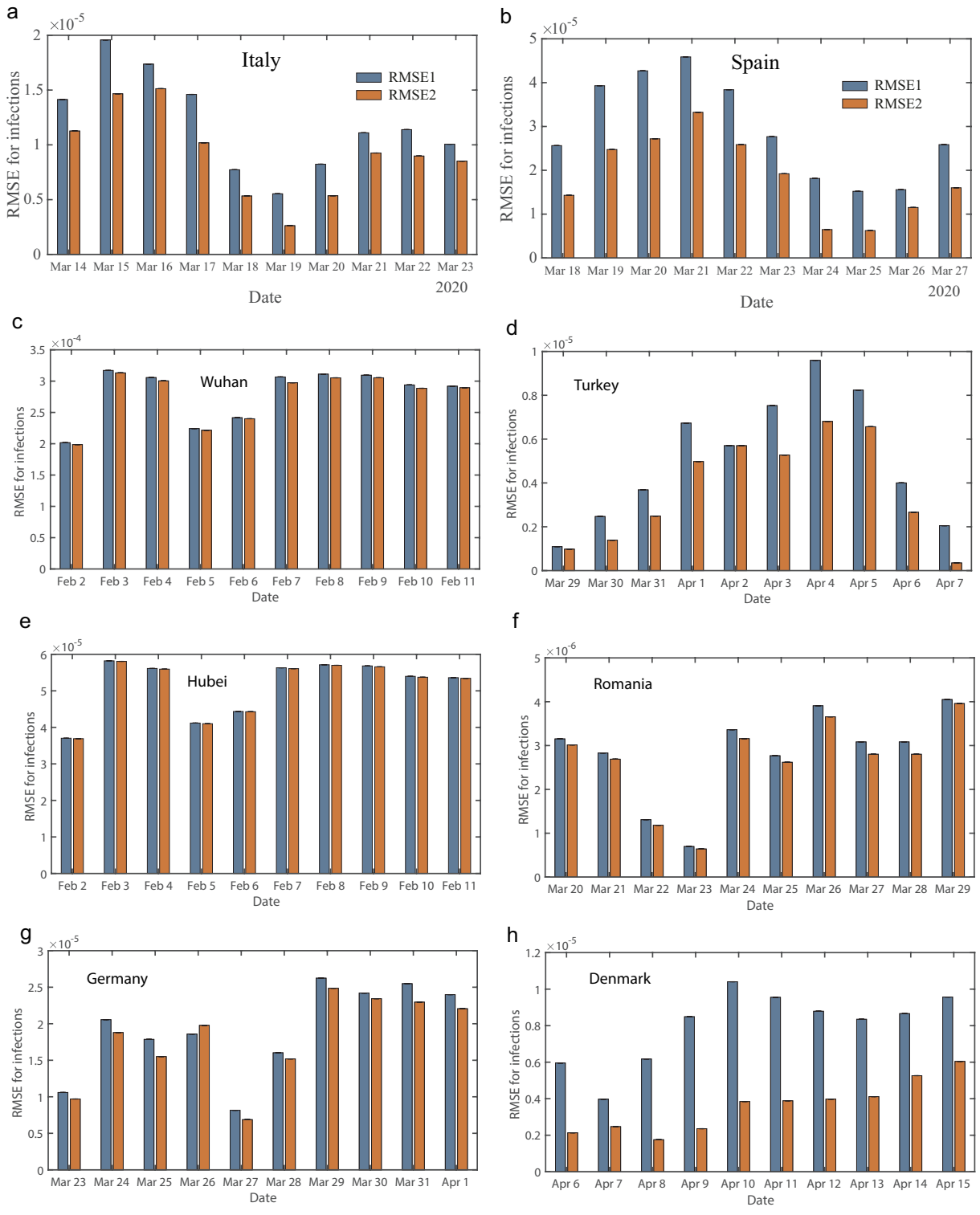

Fig. S5. Forecasts of the COVID-19 pandemic using the SIRD model with and without accounting for reporting delays. Each panel shows the RMSE between the forecast results and the real reported data when we forecast the future prevalence on different dates.

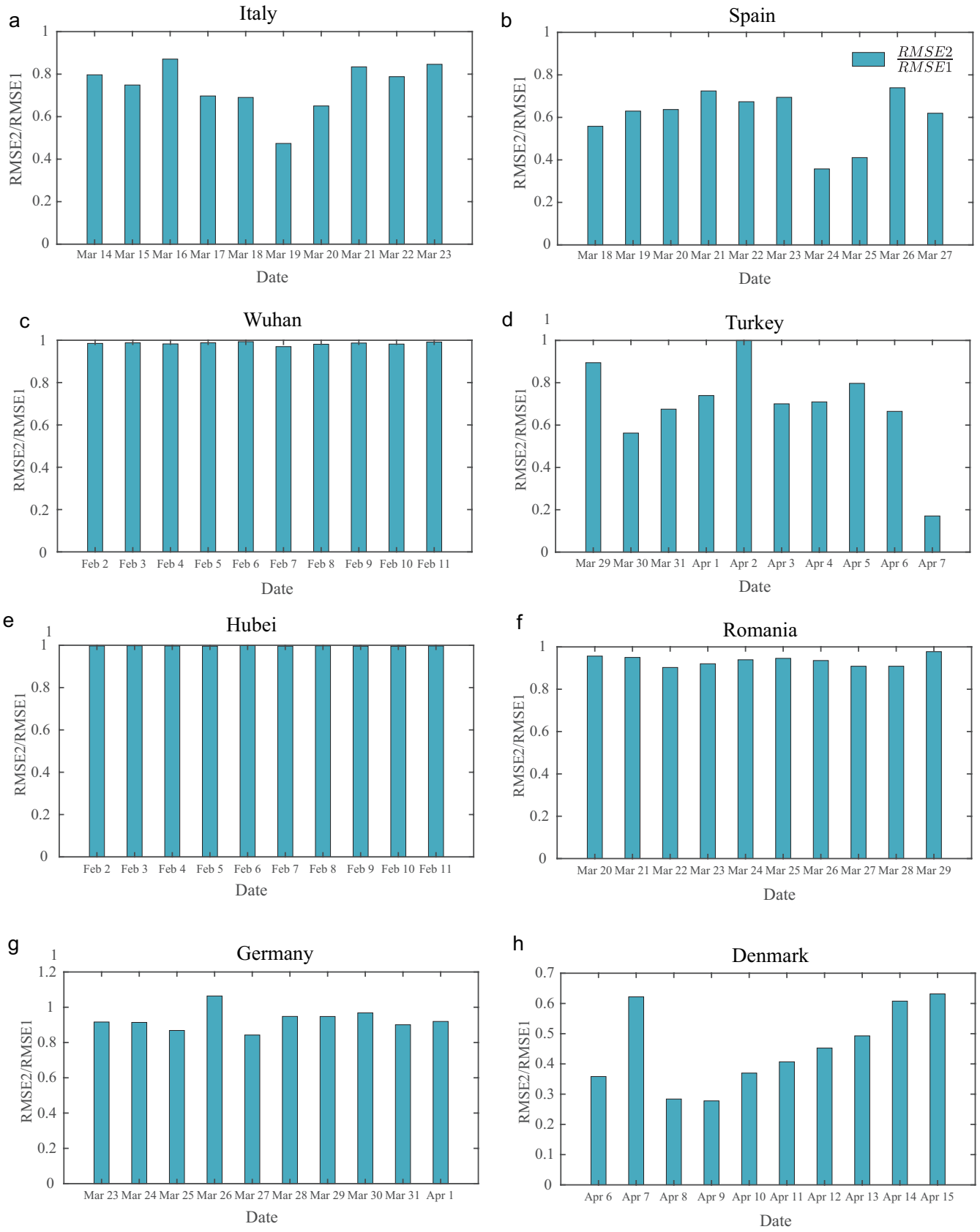

Fig. S6. The ratio between RMSE2 and RMSE1 for all 8 regions. Note that the forecast improvements are more significant for Spain, Italy, Turkey and Denmark than for the other regions.