medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Data rectification to account for delays in reporting disease incidence with an application to forecasting COVID-19 cases

Yunus A. Abdulhameed[1], Samuel Roberts[1], Jacob B. Aguilar[2], James Kercheville[1], Juan B. Gutierrez[1*].

[1] Department of Mathematics, University of Texas at San Antonio.

[2] Department of Mathematics, Saint Leo University.

*To whom correspondence should be addressed: E-mail: juan.gutierrez3@utsa.edu.

Effective monitoring of infectious disease incidence remains a major challenge to public health. Difficulties in estimating the trends in disease incidence arise mainly from the time delay between case diagnosis and the reporting of cases to public health databases. However, predictive models usually assume that public data sets faithfully reflect the state of disease transmission. In this paper, we study the effect of delayed case reporting by comparing data reported by the Johns Hopkins Coronavirus Resource Center (CRC) with that of the raw clinical data collected from the San Antonio Metro Health District (SAMHD), San Antonio, Texas. An insight on the subtle effect that such reporting errors potentially have on predictive modeling is presented. We use an exponential distribution model for the regression analysis of the reporting delay. The proposed model for correcting reporting delays was applied to our

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

recently developed SEYAR (Susceptible, Exposed, Symptomatic, Asymptomatic, Recovered) dynamical model for COVID-19 transmission dynamics. Employing data from SAMHD, we demonstrate that the forecasting ability of the SEYAR model is substantially improved when the rectified reporting obtained from our proposed model is utilized. The methods and findings demonstrated in this work have ample applicability in the forecasting of infectious disease outbreaks. Our findings suggest that failure to consider reporting delays in surveillance data can significantly alter forecasts.

1 Background

In December of 2019, a novel coronavirus (SARS-CoV-2) was first reported in the City of Wuhan, Hubei Province, China. On January 30th, 2020, the SARS-CoV-2 outbreak was declared a public health emergency of international concern. On March 11th, 2020 the World Health Organization declared a global pandemic (*43*).

Early in the pandemic, two resources rose to prominence as sources of data. The Coronavirus Resource Center (CRC) at Johns Hopkins University (*11*), and The New York Times (NYT) coronavirus database (*50*). An important characteristic of these privately-maintained public databases is that after data is initially entered, it is usually not updated.

Various models have been used during the COVID-19 outbreak. Applications of these models include to inform public health policies (*13,36*), to assess the impact of government interventions (*5, 31, 45*), to project hospital utilization (*38*), and to assess disease transmission dynamics (*33*). The outcomes of these models rely on three important factors: (i) the quality of the case data used to calibrate the model, (ii) the validity of parametric assumptions, and (iii) the number of parameters. The current focus on improving predictive

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

case modeling has been centered around the limitations that arise from poor parametric assumptions, and the large number of parameters used in some models. As a result, the problem of poor data quality has received comparatively little attention.

Quality epidemiological data is central to infectious disease surveillance and modeling. From a public health viewpoint, accurately monitored data is a crucial resource for understanding the true extent of population-level disease progression during an epidemic event (*17*). This knowledge can then be used to support critical decision making by public health authorities. Previous epidemics such as Zika, Ebola and Swine flu have revealed the usefulness of accurate epidemiological data for emergency preparedness (*9*), vaccine distribution (*22*), and planning for the future demand of critical infrastructure (*30, 41*). From an infectious disease modelling viewpoint, quality data is critical to ensuring accurate epidemic forecasting (*39*). Furthermore, the accuracy of epidemiological parameters required by compartmentalized models (*28*), agent based models (*8, 15, 24*), and/or training and validation of machine learning and statistical models (*21, 48, 52*) is directly related to the quality of input data used to estimate these parameters.

Data quality is affected by case tracing and delays in reporting. For case counts, the *epidemiological event date*, E, is often defined in practice as: (i) E = Date of onset of symptoms, (ii) if date of onset is not available, then E = date of sample collection, (iii) if date of sample collection is not available, then E = date of lab report, and (iv) if date of lab report is not available, then E = date entered into database.

The epidemiological event date changes continuously as case trace investigations take place, which can last a variable number of days depending on the operations of the municipality collecting the data. As a result, the case counts for past dates become updated with the results obtained from the case trace investigations, with older unresolved cases being dropped from the tracing process. Another source of variation in case reporting

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

is the generalized inherent delay due to the time it takes for samples to be analyzed, the time to register an identified case into a database, and overburdened surveillance systems sometimes operating with a throughput lower than the influx of new cases. It is important to note that the CRC and NYT databases do not change the data once it is reported. This raises concerns in relation to the accuracy of predictive models since many modelers do not have access to raw case tracing data.

The challenge of reporting delays was first studied by Harris (*26*) who contrived the problem as a partial multinomial distribution. Thereafter, various statistical methods (*6, 49, 55*) have been employed to address the problem of delays from infection to symptom onset, as well as the subsequent delays that arise before the data is entered into surveillance databases. These techniques broadly fall into parametric and non-parametric methods. The parametric approach accounts for reporting delay based on the assumption that delayed case data belongs to a parametric family of probability distributions. A number of parametric-based methods that model the reporting delay in previous disease outbreaks (*6, 10, 19, 27*), and in the COVID-19 pandemic (*1, 37*) have been reported.

The non-parametric approach provides an estimate of the reporting delay distribution without making any parametric assumption about the form of the underlying delay distribution. There are two common computational methods employed for finding non-parametric estimates of the reporting delay distribution. The first is based on the generalized linear model (e.g. Poisson regression or the non-parametric back-propagation of delayed cases) which requires cross-classification of reported cases by calendar time of diagnosis and reporting delays (*6, 55*). The back-propagation process uses a diagnosis distribution to estimate the number of case counts for previous time periods (*4, 32, 34*), and has recently been used to assess infection incidence of COVID-19 (*37*). The second method employs a survival analysis approach which involves expressing the delay distri-

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

bution as a product of conditional probabilities, from which an estimate for the reporting delays is obtained.

Another approach used to account for delays in case data produces real-time estimates for the current number of confirmed infections while correcting for underreporting. This new technique is referred to as 'nowcasting' (*18, 20, 42, 51, 54*). Depending on the assumptions it is premised on, this method could be considered either a parametric or a non-parametric approach. For example, nowcasting procedures based on the non-parametric approach have been used to assess the Shiga toxin–producing E. coli (STEC) O104:H4 outbreak in Germany (*29*). On the other hand, to study influenza A/H1N1 during the 2009 pandemic, a nowcasting procedure that involved the parametric method was utilized (*12*). In addition, nowcasting has been used for real-time COVID-19 tracking (*23, 25, 47*).

Data assimilation techniques have also been used (*14, 16, 40*) to study reporting delays. These methods begin with a wide prior distribution for the model parameters from which a posterior estimation of parameters leads to model predictions that closely agree with the observations. An interesting example of data assimilation is Abott et al. (*1*), since there is explicit consideration of reporting delay. In all these methods, model updates occur when new observations become available.

In this present study, we consider an approach that is notably different from nowcasting and data assimilation. It is informed by the comparison of the totality of records of daily cases from public databases against the real-time number of hospitalizations and cases. The COVID-19 data used here is for the City of San Antonio, Texas, the seventh largest city in the US, with a population of 1.5 million people. It is the most visited city in Texas, and the 17th most visited city in the US, attracting 37 million visitors in 2019 (*3*). Given the high density of visitors year-round, the city has the potential to become an epicenter of transmission during a pandemic. The City of San Antonio entered a partnership with

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Bexar County to create the San Antonio Metro Health District (SAMHD); this entity is in charge of collecting case data for infectious diseases.

To better account for delays in reporting COVID-19 case-data, and to accurately forecast the number of confirmed cases, we adopted the following approach. First, we analysed the complete set of epidemiological event dates for confirmed COVID-19 cases and developed an algorithm that rectifies delays in public epidemiological data. Then, we incorporated the proposed rectification algorithm into our compartmentalized SEYAR model to project the number of cases in Bexar County, Texas, whilst taking into account reporting delays.

2 Methods

2.1 Epidemiological data extraction

We employed two different data sets which were obtained from two different sources. The first being the COVID-19 case-data retrieved from the CRC. The second was the data regarding the epidemiological date for COVID-19 cases acquired from the SAMHD between April 4, 2020 through June 28, 2020. The epidemiological event dates received daily from the SAMHD include; (i) illness onset, (ii) sample collection, (iii) test result, and (iv) case entered into the database. It is usually not the case that information is available for every epidemiological date mentioned ((i)-(iv)); but at least one of these dates is available for each confirmed case.

2.2 Extracting the delay distribution from surveillance data

Let a_{jk} be the data reported on day k into the SAMHD registry, for each day j subject to $1 \leq j \leq k$. Let b_{jk} be the 'stable' data where 'stable' means that as time progresses, the number of cases for day b_{jk} does not change. The residue $r_{jk} = a_{jk} - b_{jk}$ is modeled

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

as an exponential distribution characterized by $y_j = pe^{qj}$ where the parameters p and q denote the correction rate.

The least-squares objective function is expressed as

$$\begin{aligned} \min_{p,q} J(p,q) &= \frac{1}{2}\sum_{j=1}(pe^{qj} - y_j)^2 \\ &= \frac{1}{2}\sum_{j=1} f_j(p,q)^2 \\ &= \frac{1}{2}F(p,q)^T F(p,q) \end{aligned} \tag{1}$$

where F is the vector-valued function $F(p,q) = (f_1(p,q) \quad f_2(p,q) \quad \ldots \quad f_i(p,q))^T$. The derivatives are made less cluttered by scaling the problem by $\frac{1}{2}$. The gradient of f is

$$\nabla f(p,q) = \frac{1}{2}\begin{pmatrix} \sum_{j=1}(pe^{qj} - y_j)e^{qj} \\ \sum_{j=1}(pe^{qj} - y_j)pje^{qj} \end{pmatrix}. \tag{2}$$

Suppose the solution of the least-squares problem is given by p_*, q_*, then $f(p_*,\mathrm{q}_*) = 0$. This implies that for all j, $f_j(p_*, q_*) = 0$, suggesting that the model is in agreement with the data with minimal error. Consequently, $F(p_*, q_*) = 0$ for $p \approx p_*$, $q \approx q_*$ which justifies that the required first-order condition is met.

After training, the obtained parameters were averaged and used to test the model. Finally, the averaged estimated parameters obtained from the model validation process were employed to estimate the delayed daily case counts for each data set reported by the CRC.

2.3 Minimizing the time lag between CRC and SAMHD data

The number of confirmed cases for Bexar County entered into the SAMHD registry was updated each day to reflect the number of delayed cases, hence providing an accurate

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

baseline. Our objective is to minimize the error between the data entered into the SAMHD and CRC registries.

Let a_{mk} and c_{nk} be the case data reported into the SAMHD and CRC registries respectively on day k, for each day m (for SAMHD data) and n (for CRC data) subject to $1 \leq m \leq k$ and $1 \leq n \leq k$ respectively. The optimal data time Δt that minimizes the time lag between these two case data is estimated by taking the difference between the dates at which the number of cases in a_{mk} and c_{nk} are equal. The optimal data time objective function is an average expressed as

$$\begin{aligned} y =& \overline{\min\left[t_{mk} - t_{nk}\right]} \\ \text{subject to:} \quad & a_{mk}(t) = c_{nk}(t). \end{aligned} \tag{3}$$

Finally, $\Delta t = \overline{y}$.

2.4 SEYAR Model Modification

Consider the *SEYAR* dynamical system (4) introduced by Aguilar et. al (*2*), which describes the dynamics of COVID-19 transmission in a human population by decomposing the total host population (N) into the following five epidemiological classes: susceptible human (S), exposed (E), symptomatic (Y), asymptomatic (A), and recovered (R).

$$\begin{cases} \dot{S} = -\left(\beta_Y \frac{Y}{N} + \beta_A \frac{A}{N}\right) SQ, \\ \dot{E} = \left(\beta_Y \frac{Y}{N} + \beta_A \frac{A}{N}\right) SQ - \gamma E, \\ \dot{Y} = \gamma(1-\alpha)E - (\delta + \lambda_{YR})\, Y, \\ \dot{A} = \gamma\alpha E - \lambda_{AR} A, \\ \dot{R} = \lambda_{AR} A + \lambda_{YR} Y. \end{cases} \tag{4}$$

Here β_Y, β_A denote the effective contact rates for symptomatic and asymptomatic carriers respectively, γ represents the latent period, λ_{YR}^{-1}, λ_{AR}^{-1} represent the infectious periods for

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the symptomatic and asymptomatic sub-populations, and α, $(1-\alpha)$ represent the probability of becoming asymptomatic and symptomatic upon infection respectively. Moreover, the risk presented at time t is represented by $Q(t) = e^{-kt}$ and is formulated under the assumption that the rate of change of risk decreases proportionally to the amount of risk present. In the presence of reporting delay, the amount of risk at any time t is proportional to the corrected number of reported COVID-19 cases, i.e. the number of corrected cases increases with an increase in the risk of infection. Thus, we arrive at the following modification of (4):

$$\begin{cases} \dot{S} = -\left(\beta_Y \frac{Y}{N} + \beta_A \frac{A}{N}\right) Spe^{j(q-k)}, \\ \dot{E} = \left(\beta_Y \frac{Y}{N} + \beta_A \frac{A}{N}\right) Spe^{j(q-k)} - \gamma E, \\ \dot{Y} = \gamma(1-\alpha)E - (\delta + \lambda_{YR})\, Y, \\ \dot{A} = \gamma\alpha E - \lambda_{AR} A, \\ \dot{R} = \lambda_{AR} A + \lambda_{YR} Y. \end{cases} \tag{5}$$

where the parameters p and q were fitted for a given value of the coefficient of risk mitigation k.

To compute the confidence intervals, a sum of 100 bootstrap replications of the case time series were used to find the possible value of the parameter that is close to a global minimum. With the parameters obtained from bootstrapping, the SEYAR model was computed using both datasets. When computing, the model considered the past two weeks of data from the most recent date of available case data.

3 Results

Figure 1 summarizes the daily number of confirmed cases in Bexar County, Texas due to the COVID-19 pandemic, as reported by both the SAMHD and CRC. The various solid curves illustrate the daily numbers of cases reported by SAMHD between May 4th 2020 through July 4th 2020. In particular, each solid curve represents a corrected

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1: Average values of parameters used to compute Figure 3. The fitted parameters were selected from the optimal fit of the model calibrated for Bexar County.

Parameter	**Description**	**Dimension**	**Value**	
$\boxed{FITTED}$				
β_Y	Effective contact rate from symptomatic to susceptible.	days^{-1}	1.17	
β_A	Effective contact rate from asymptomatic to susceptible.	days^{-1}	1.16	
k	Risk mitigation coefficient	n/a	0.03	
n	Days prior, patient zero	n/a	25	
Parameter	**Description**	**Dimension**	**Mean**	**Variance**
$\boxed{BIOLOGICAL}$				
γ^{-1}	Mean latent period.	days	2.99	1.36
α	Probability of becoming asymptomatic upon infection.	n/a	0.21	1.48
λ_{YR}^{-1}	Mean symptomatic infectious period.	days	13.5	31.8
λ_{AR}^{-1}	Mean asymptomatic infectious period.	days	11.2	16
δ	Disease-induced death rate.	days^{-1}	0.026	0.094

version of the number of cases reported in the past. The dashed curve represents the data reported by the CRC between May 4th 2020 through July 4th 2020. In contrast to the SAMHD data sets, the CRC data for the number of cases reported in the past are not corrected. Moreover, the time lag between the solid and dashed curves was computationally estimated at 8 days.

In Figure 2, the CRC data sets reported between May 4th 2020 and July 4th 2020 were rectified using our proposed data rectification algorithm which was subsequently compared with the baseline SAMHD data set reported on July 4th and the individual unrectified CRC data sets. The Figure indicates congruence between the data transmitted on the same date by the SAMHD, and the rectified CRC data. Additionally, it can be observed that the SAMHD and the CRC reported different numbers of cases for the

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

majority of dates. A significant increase in reporting delay of daily case counts is noticed from approximately the beginning of June and through the conclusion of this study.

Taking into account the parameters reported in Table 1, and the five major government mitigation policies that were enacted in Bexar county between March 2020 and July 2020, Figure 3 summarizes an implementation of a SEYAR model calibrated with three sets of data: the time series obtained from CRC, the rectified CRC data, and the data collected from SAMHD. Figure 3 (a) shows a comparison between the average SEYAR model output calibrated using the un-rectified CRC data set reported on June 30th 2020 and the data retrospectively obtained for the month of October 2020. Figure 3 (b) shows a comparison between the average SEYAR model output calibrated using the rectified CRC data set reported on June 30th 2020 and the data retrospectively obtained for the month of October 2020. Figure 3 (c) shows a comparison between the average SEYAR model output calibrated using the SAMHD data set reported on June 30th 2020 and the data retrospectively obtained for the month of October 2020.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

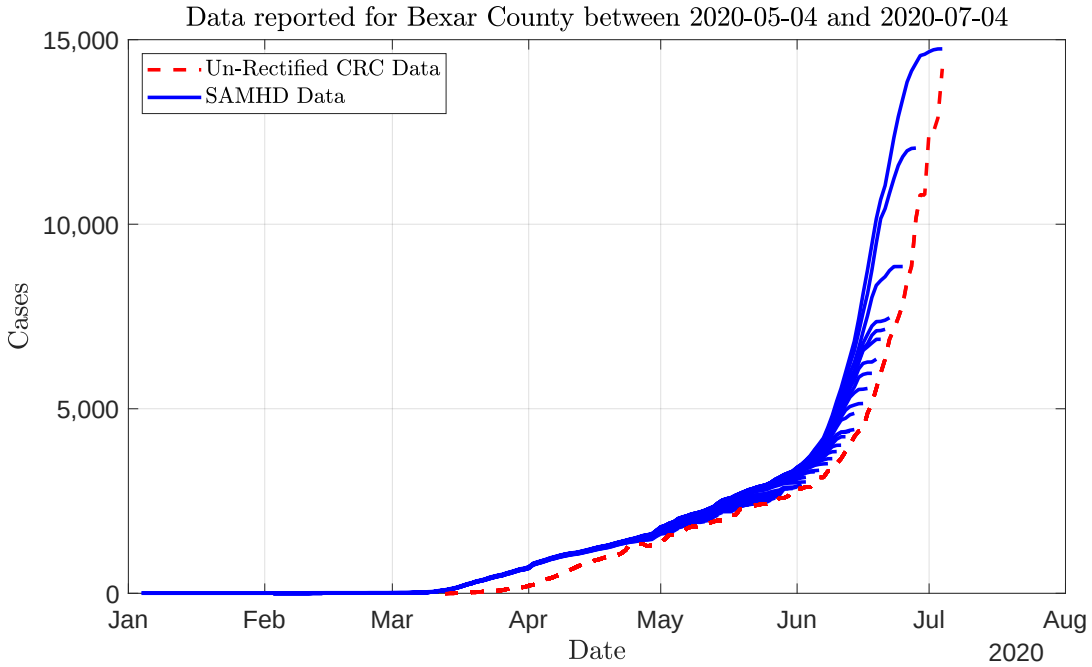

Figure 1: Daily number of confirmed COVID-19 cases in Bexar County, Texas reported for the period between May 4th 2020 through July 4th 2020. Comparison of daily case counts reported by the San Antonio Metro Health District (SAMHD) and the Johns Hopkins Coronavirus Resource Center (CRC). The solid curves represent individual SAMHD data sets across this time period for which previously reported data sets are corrected. The dashed curve refers to the CRC data set for which no corrections are made in the data sets reported in the past.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

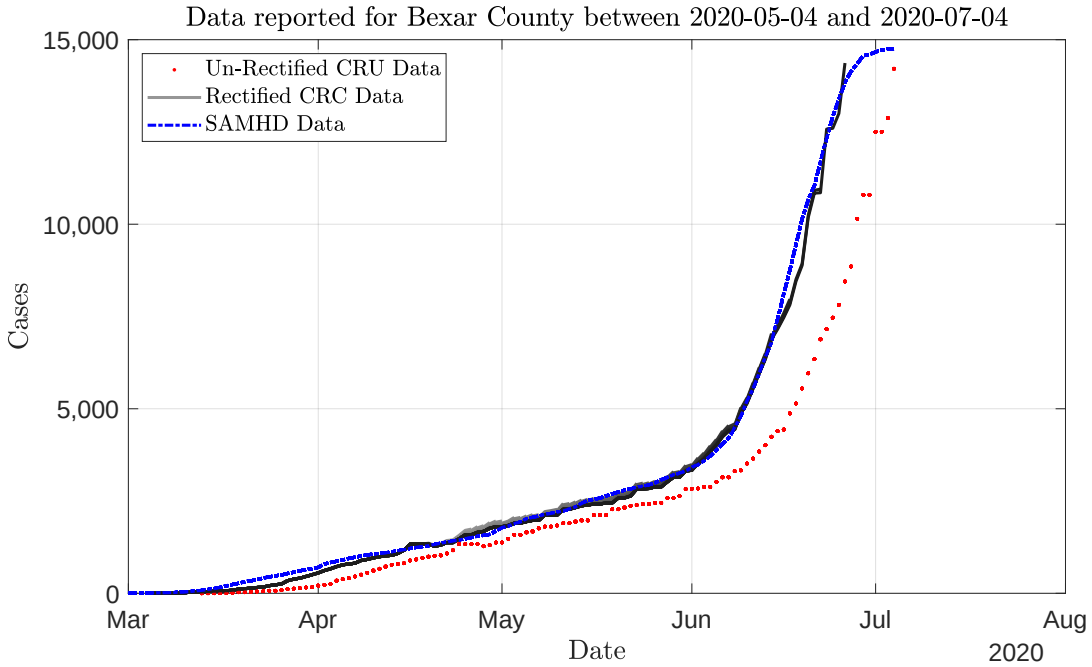

Figure 2: Rectification for individual CRC data sets reported between May 4th 2020 through July 4th, 2020. The solid curve indicate the rectified daily confirmed cases that were reported during the pandemic. The dash-dotted curve represents our baseline data, i.e the data set reported by the SAMHD on July 4th 2020. The dotted curve represents the un-rectified CRC data between May 4th 2020 through July 4th 2020.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

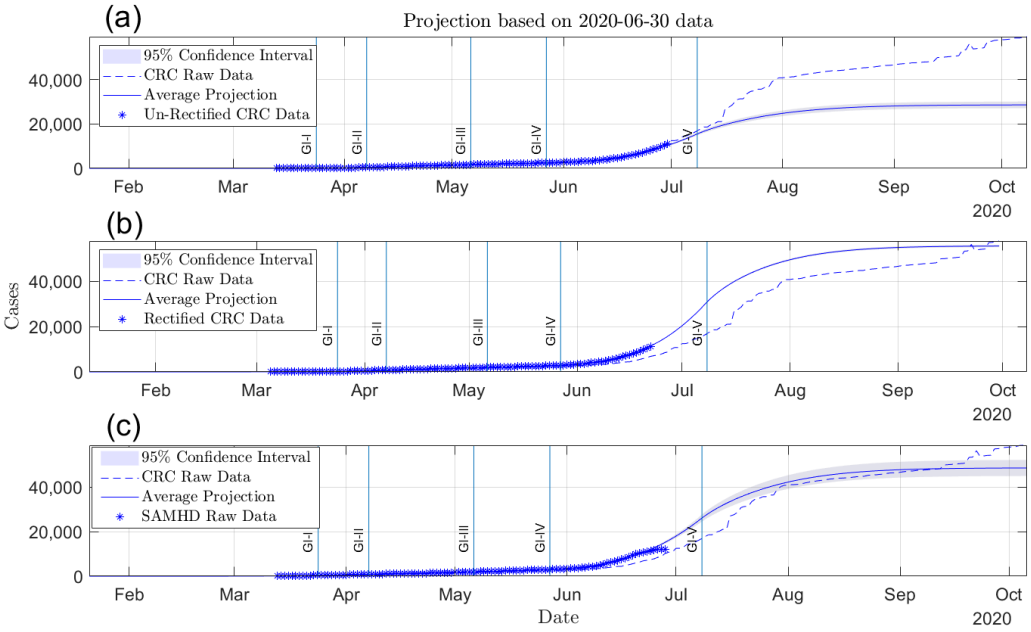

Figure 3: Projections of COVID-19 cases in Bexar County from July 2020 to October 2020, using a SEYAR model calibrated with three data sets: (a) the time series obtained from CRC, (b) the rectified CRC data, and (c) the data collected from SAMHD. The projected number of cases with 95% confidence interval for each of the data sets were obtained under the conditions that applied on July 4th 2020. The blue-dashed line (CRC raw data) denotes the un-rectified daily number of cases reported by the Johns Hopkins Coronavirus Resource Center. The solid vertical lines indicate the timings of various government intervention (GI) strategies.

4 Discussion

By analyzing the COVID-19 case count data for Bexar County, Texas, obtained from two different sources –the SAMHD and CRC, we have been able to study how reporting delays in CRC case-data could affect the predictability of trends in the number of daily confirmed cases, as well as the quality of data on which predictive models are calibrated. Accounting for reporting delays in epidemic data enhances the ability to gauge the actual daily case counts for forecasting reliable trends in the number of confirmed cases. This

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

is supported by the findings from the works of Brookmeyer and Liao (*7*), White and Pagano (*53*).

Many of the aforementioned studies on reporting delay rely on an underlying assumed probability density function such as Poisson, Binomial etc. Whilst there are theoretical basis for making such assumptions, unless the chosen distribution is measured from sample data, it remains an assumption which has the potential for introducing a bias. Instead of trying to fit parameters for assumed distributions that make a model fit the data, this present study fits a regression from the complete analysed case tracing data, making the minimum number of assumptions possible. By considering past dates of the pandemic, we can retrospectively evaluate the accuracy of our rectification and forecasting estimates for the number of confirmed cases in Bexar County.

The data pattern that emerged in San Antonio is captured in Figure 1. The solid curves represent cases reported by SAMHD in multiple days. The dashed curve represents data reported by CRC, a source that never updates its initial data entries. It is evident from Figure 1 that case records are constantly corrected at SAMHD by case tracing, and they tend to mimic daily CRC counts with a delay. If data obtained from public databases is substantially different from the true count of cases, then this begs the question of how to create accurate predictive models when most modelers have access only to public databases.

Under-reporting of confirmed cases poses a potentially major challenge in COVID-19 surveillance. Early investigations suggested that most cases are not reported to the Centers for Disease Control (*44*). This justifies our analysis of the cases reported by the CRC for Bexar County (Figure 2), which indicates an incomplete number of daily cases reported. More recent findings indicated a detection rate of only 1–2 % of total actual COVID-19 cases (*35*). As revealed by our analysis of reporting delays, the rate of

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

unreported cases may vary depending on the time lapse between identification of a case and reporting to public registries.

The significant differences observed in the daily number of confirmed cases reported to the CRC is related to the effect of variation in the chosen epidemiological event date caused by case tracing. As a consequence of overwhelmed surveillance systems during the pandemic, there was an usually long turnaround time between the date of onset of illness, date of diagnosis, date of laboratory sample collection, laboratory test result date, and the date that a confirmed case is entered into an official database. These delays in processing lead to imprecise day-to-day case reporting. These challenges have also been shown to introduce a bias in the estimate of case fatality ratio (*46*).

The minimization of the time lag from the reporting of cases and the subsequent adjustment of daily case counts using our proposed method produced a rectification of the public CRC data that approximated data in the official database after case tracing corrections. In terms of rectification of delayed cases, the exponential distribution has provided an accurate model for adjusting delay in case reporting. The agreement between the SAMHD data and rectified CRC data (as shown in Figure 2) indicates the reliability of the proposed rectification method.

When using an unrectified CRC data set for the SEYAR model calibration, projections were significantly lower as compared to those obtained using the CRC data set, as shown in Figure 3 (a). On the other hand when using a rectified data set for calibration, projections were slightly higher than the CRC data, as shown in Figure 3(b). These findings suggest that forecasts obtained following model calibration with rectified data are less biased. Using the SAMHD data for calibration, the model's average projected cases for July through mid-September 2020 were nearly in agreement with the CRC data (Figure 3 (c)).

It is worth noting that the distribution used in this study is measured from case

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

data that were recorded during the early phase of the COVID-19 pandemic. Making the application of our proposed algorithm more suitable during the exponential growth phases of an outbreak.

Some possible limitations of the rectification algorithm deserve comment. First, the SAMHD data was available until July 7, 2020. After that date, changes in the reporting system made case tracing adjustments unavailable. Thereafter, the only baseline for comparison became CRC. Second, the measure of the delay distribution was based on data for Bexar County, Texas, (COVID-19 epidemiological event dates of confirmed cases) obtained from the San Antonio Metro Health District; based on our data, the optimal time that minimizes the reporting delay in CRC data was computationally estimated at 8 days. If there is evidence of a significantly higher delay in reporting in a different geographical region, the optimal data time should be measured for that location to reduce bias in the rectification.

The effect of reporting delays in data used for forecasting is often overlooked by many predictive models. Those models which utilize daily case counts often assume that the public case data are a faithful account of reality, which is hardly the case in practice as reporting delays in epidemic data remain inevitable. It is expected that the use of our methodological approach for rectifying and minimizing reporting delay when implementing prediction models could significantly improve the predictability of an epidemic.

5 Conclusion

To analyze the delays in case reporting to public health databases, we utilized both surveillance data and data obtained from public databases. We developed a method that rectifies such delays in public epidemiological data. Unlike the approaches reported in the literature, the rectification algorithm proposed here is premised on an exponential

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

distribution that is measured from case tracing data with minimal assumptions. The method was shown to reliably account for the delays in reporting. The rectification of reporting delays in conjunction with our SEYAR prediction model seems promising for forecasting the actual number of daily cases. We stress the importance of capturing and publishing daily case counts as the data changes over time. This is particularly important as public databases such as the CRC do not currently update their daily case count records. This study has potential importance in terms of assessing the severity of a pandemic, particularly during its early stages. The results presented here also have the potential to aid in assessing the impact of possible control measures.

References

1. Sam Abbott, Joel Hellewell, Robin N Thompson, Katharine Sherratt, Hamish P Gibbs, Nikos I Bosse, James D Munday, Sophie Meakin, Emma L Doughty, June Young Chun, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research*, 5(112):112, 2020.

2. Jacob B Aguilar, Jeremy Samuel Faust, Lauren M Westafer, and Juan B Gutierrez. Modeling the impact of asymptomatic carriers on COVID-19 transmission dynamics during lockdown. *MedRxiv*, 2020.

3. Visit San Antonio. Economic Impact of San Antonio's Tourism Industry Rises to $15.2 Billion, 2020. https://meetings.visitsanantonio.com/economic-impact-of-san-antonios-tourism-industry-rises-to-15-2-billion/. Accessed on April 1, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

4. Niels G Becker, Lyndsey F Watson, and John B Carlin. A method of non-parametric back-projection and its application to AIDS data. *Statistics in Medicine*, 10(10):1527–1542, 1991.

5. David Berger, Kyle Herkenhoff, Chengdai Huang, and Simon Mongey. Testing and reopening in an SEIR model. *Review of Economic Dynamics*, 2020.

6. Ron Brookmeyer and Anne Damiano. Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine*, 8(1):23–34, 1989.

7. Ron Brookmeyer and Mitchell H Gail. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association*, 83(402):301–308, 1988.

8. Dennis L Chao, Scott B Halstead, M Elizabeth Halloran, and Ira M Longini Jr. Controlling dengue with vaccines in Thailand. *PLoS Neglected Tropical Diseases*, 6(10):e1876, 2012.

9. Jean-Paul Chretien, Caitlin M Rivers, and Michael A Johansson. Make data sharing routine to prepare for public health emergencies. *PLoS Medicine*, 13(8):e1002109, 2016.

10. Stephen R Cole, Haitao Chu, and Sander Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American Journal of Epidemiology*, 179(2):252–260, 2014.

11. Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

12. Tjibbe Donker, Michiel van Boven, W Marijn van Ballegooijen, Tessa M van't Klooster, Cornelia C Wielders, and Jacco Wallinga. Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands. *European Journal of Epidemiology*, 26(3):195–201, 2011.

13. Dileepa Senajith Ediriweera, Nilanthi Renuka De Silva, Gathsaurie Neelika Malavige, and Hithanadura Janaka De Silva. An epidemiological model to aid decision-making for COVID-19 control in Sri Lanka. *PLoS One*, 15(8):e0238340, 2020.

14. Ralf Engbert, Maximilian M Rabe, Reinhold Kliegl, and Sebastian Reich. Sequential data assimilation of the stochastic seir epidemic model for regional covid-19 dynamics. *Bulletin of mathematical biology*, 83(1):1–16, 2021.

15. Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

16. Geir Evensen, Javier Amezcua, Marc Bocquet, Alberto Carrassi, Alban Farchi, Alison Fowler, Peter Houtekamer, Christopher KRT Jones, Rafael de Moraes, Manuel Pulido, et al. An international assessment of the covid-19 pandemic using ensemble data assimilation. *medRxiv*, 2020.

17. Geoffrey Fairchild, Byron Tasseff, Hari Khalsa, Nicholas Generous, Ashlynn R Daughton, Nileena Velappan, Reid Priedhorsky, and Alina Deshpande. Epidemiological data challenges: planning for a more robust future through data standards. *Frontiers in Public Health*, 6:336, 2018.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

18. CP Farrington, Nick J Andrews, AD Beale, and MA Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(3):547–563, 1996.

19. Mitchell H Gail and Ron Brookmeyer. Methods for projecting course of acquired immunodeficiency syndrome epidemic1. *JNCI: Journal of the National Cancer Institute*, 80(12):900–911, 1988.

20. Tini Garske, Judith Legrand, Christl A Donnelly, Helen Ward, Simon Cauchemez, Christophe Fraser, Neil M Ferguson, and Azra C Ghani. Assessing the severity of the novel influenza A/H1N1 pandemic. *Bmj*, 339, 2009.

21. Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

22. Edward Goldstein, J Wallinga, and Marc Lipsitch. Vaccine allocation in a declining epidemic. *Journal of The Royal Society Interface*, 9(76):2798–2803, 2012.

23. Sharon K Greene, Sarah F McGough, Gretchen M Culp, Laura E Graf, Marc Lipsitch, Nicolas A Menzies, and Rebecca Kahn. Nowcasting for real-time COVID-19 tracking in New York City: An evaluation using reportable disease data from early in the pandemic. *JMIR Public Health and Surveillance*, 7(1):e25538, 2021.

24. John J Grefenstette, Shawn T Brown, Roni Rosenfeld, Jay DePasse, Nathan TB Stone, Phillip C Cooley, William D Wheaton, Alona Fyshe, David D Galloway, Anuroop Sriram, et al. Fred (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health*.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

25. Felix Günther, Andreas Bender, Katharina Katz, Helmut Küchenhoff, and Michael Höhle. Nowcasting the covid-19 pandemic in bavaria. *Biometrical Journal*, 63(3):490–502, 2021.

26. Jeffrey E Harris. Delay in reporting acquired immune deficiency syndrome (AIDS). *NBER Working Paper*, (w2278), 1987.

27. Jeffrey E Harris. Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association*, 85(412):915–924, 1990.

28. Herbert W Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

29. Michael Höhle and Matthias an der Heiden. Bayesian nowcasting during the STEC o104: H4 outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002, 2014.

30. Susy Hota, Elchanan Fried, Lisa Burry, Thomas E Stewart, and Michael D Christian. Preparing your intensive care unit for the second wave of H1N1 and future surges. *Critical Care Medicine*, 38:e110–e119, 2010.

31. Daniel CP Jorge, Moreno S Rodrigues, Mateus S Silva, Luciana L Cardim, Nívea B da Silva, Ismael H Silveira, Vivian AF Silva, Felipe AC Pereira, Arthur R de Azevedo, Alan AS Amad, et al. Assessing the nationwide impact of COVID-19 mitigation policies on the transmission rate of SARS-CoV-2 in Brazil. *Epidemics*, 35:100465, 2021.

32. JD Kalbfleisch and Jerald F Lawless. Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association*, 84(406):360–372, 1989.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

33. Stephen M Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H Grad, and Marc Lipsitch. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*, 368(6493):860–868, 2020.

34. S Wt Lagakos, LM Barraj, and V de Gruttola. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75(3):515–523, 1988.

35. Hien Lau, Tanja Khosrawipour, Piotr Kocbach, Hirohito Ichii, Jacek Bania, and Veria Khosrawipour. Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. *Pulmonology*, 27(2):110–115, 2021.

36. Kathy Leung, Joseph T Wu, Di Liu, and Gabriel M Leung. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *The Lancet*, 395(10233):1382–1393, 2020.

37. IC Marschner. Back-projection of COVID-19 diagnosis counts to assess infection incidence and control measures: analysis of Australian data. *Epidemiology & Infection*, 148, 2020.

38. Seyed M Moghadas, Affan Shoukat, Meagan C Fitzpatrick, Chad R Wells, Pratha Sah, Abhishek Pandey, Jeffrey D Sachs, Zheng Wang, Lauren A Meyers, Burton H Singer, et al. Projecting hospital utilization during the COVID-19 outbreaks in the United States. *Proceedings of the National Academy of Sciences*, 117(16):9122–9126, 2020.

39. Kelly R Moran, Geoffrey Fairchild, Nicholas Generous, Kyle Hickmann, Dave Osthus, Reid Priedhorsky, James Hyman, and Sara Y Del Valle. Epidemic forecasting is messier than weather forecasting: the role of human behavior and internet data

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

streams in epidemic forecast. *The Journal of Infectious Diseases*, 214(suppl_4):S404–S408, 2016.

40. Philip Nadler, Shuo Wang, Rossella Arcucci, Xian Yang, and Yike Guo. An epidemiological modelling approach for covid-19 via data assimilation. *European Journal of Epidemiology*, 35(8):749–761, 2020.

41. Raoul E Nap, Maarten PHM Andriessen, Nico EL Meessen, and Tjip S Van der Werf. Pandemic influenza and hospital resources. *Emerging Infectious Diseases*, 13(11):1714, 2007.

42. A Nicoll, A Ammon, A Amato, B Ciancio, P Zucs, I Devaux, F Plata, A Mazick, K Mølbak, T Asikainen, et al. Experience and lessons from surveillance and studies of the 2009 pandemic in Europe. *Public Health*, 124(1):14–23, 2010.

43. World Health Organization et al. WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020, 2020.

44. Marcelo Freitas do Prado, Bianca Brandão de Paula Antunes, Leonardo dos Santos Lourenço Bastos, Igor Tona Peres, Amanda de Araújo Batista da Silva, Leila Figueiredo Dantas, Fernanda Araújo Baião, Paula Maçaira, Silvio Hamacher, and Fernando Augusto Bozza. Analysis of COVID-19 under-reporting in Brazil. *Revista Brasileira de Terapia Intensiva*, 32:224–228, 2020.

45. Benjamin Roche, Andres Garchitorena, and David Roiz. The impact of lockdown strategies targeting age groups on the burden of COVID-19 in France. *Epidemics*, 33:100424, 2020.

46. Timothy W Russell, Joel Hellewell, Sam Abbott, N Golding, H Gibbs, CI Jarvis, K van Zandvoort, S Flasche, R Eggo, WJ Edmunds, et al. Using a delay-adjusted

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

case fatality ratio to estimate under-reporting. *Centre for Mathematical Modeling of Infectious Diseases Repository*, pages 1–6, 2020.

47. Marc Schneble, Giacomo De Nicola, Göran Kauermann, and Ursula Berger. Nowcasting fatal covid-19 infections on a regional level in germany. *Biometrical Journal*, 63(3):471–489, 2021.

48. Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*, 4(1):1–10, 2013.

49. Donna F Stroup, G David Williamson, Joy L Herndon, and John M Karon. Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine*, 8(3):323–329, 1989.

50. New York Times. Coronavirus in the U.S.: Latest Map and Case Count, 2021. https://www.nytimes.com/interactive/2021/us/covid-cases.html. Accessed on September 1, 2021.

51. Jan van de Kassteele, Paul HC Eilers, and Jacco Wallinga. Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained p-spline smoothing. *Epidemiology (Cambridge, Mass.)*, 30(5):737, 2019.

52. Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006, 2003.

53. Laura F White and Marcello Pagano. Reporting errors in infectious disease outbreaks, with an application to Pandemic Influenza A/H1N1. *Epidemiologic Perspectives & Innovations*, 7(1):1–12, 2010.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305398; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

54. Joseph T Wu, Kathy Leung, and Gabriel M Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in wuhan, china: a modelling study. *The Lancet*, 395(10225):689–697, 2020.

55. Scott L Zeger, Lai-Chu See, and Peter J Diggle. Statistical methods for monitoring the AIDS epidemic. *Statistics in Medicine*, 8(1):3–21, 1989.