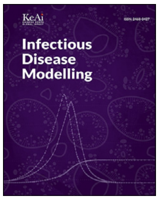
Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm



A simple modification to the classical SIR model to estimate the proportion of under-reported infections using case studies in flu and COVID-19

Leonid Kalachev [a, b, *], Jon Graham [a, b], Erin L. Landguth [b]

[a] Mathematical Sciences, University of Montana, Missoula, USA
[b] Center for Population Health Research, School of Public and Community Health Sciences, University of Montana, Missoula, USA

ARTICLE INFO

Article history:
Received 21 March 2024
Received in revised form 7 May 2024
Accepted 14 June 2024
Available online 16 June 2024
Handling Editor: Daihai He

Keywords:
Modeling of epidemics
SIR type models
flu and COVID-19
Under-reporting parameter
Proportion of reported disease cases

ABSTRACT

Background: Under-reporting and, thus, uncertainty around the true incidence of health events is common in all public health reporting systems. While the problem of under-reporting is acknowledged in epidemiology, the guidance and methods available for assessing and correcting the resulting bias are obscure.

Objective: We aim to design a simple modification to the Susceptible – Infected – Removed (SIR) model for estimating the fraction or proportion of reported infection cases.

Methods: The suggested modification involves rescaling of the classical SIR model producing its mathematically equivalent version with explicit dependence on the reporting parameter (true proportion of cases reported). We justify the rescaling using the phase plane analysis of the SIR model system and show how this rescaling parameter can be estimated from the data along with the other model parameters.

Results: We demonstrate how the proposed method is cross-validated using simulated data with known disease cases and then apply it to two empirical reported data sets to estimate the fraction of reported cases in Missoula County, Montana, USA, using: (1) flu data for 2016–2017 and (2) COVID-19 data for fall of 2020.

Conclusions: We establish with the simulated and COVID-19 data that when most of the disease cases are presumed reported, the value of the additional reporting parameter in the modified SIR model is close or equal to one, so that the original SIR model is appropriate for data analysis. Conversely, the flu example shows that when the reporting parameter is close to zero, the original SIR model is not accurately estimating the usual rate parameters, and the re-scaled SIR model should be used. This research demonstrates the role of under-reporting of disease data and the importance of accounting for under-reporting when modeling simulated, endemic, and pandemic disease data. Correctly reporting the “true” number of disease cases will have downstream impacts on predictions of disease dynamics. A simple parameter adjustment to the SIR modeling framework can help alleviate bias and uncertainty around crucial epidemiological metrics (e.g.: basic disease reproduction number) and public health decision making.

* Corresponding author. University of Montana, 32 Campus Drive, Missoula, MT, 59812. USA.
E-mail addresses: kalachev@mso.umt.edu (L. Kalachev), erin.landguth@mso.umt.edu (E.L. Landguth).

Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2024.06.002

1. Introduction

Accurately capturing infectious disease incidence data is vital for monitoring trends and disease outbreaks as well as providing trusted predictions. Limitations exist in using health event data from most reporting systems, and most are known to have under-reporting and, thus, uncertainty around the "true" disease incidence (Keramarou & Evans, 2012; Tam et al., 2012; Wu et al., 2020). Under-reporting in the health care system can reflect many possibilities of why reporting systems are not able to accurately capture all infections within a population. For example, not all people who are infected with a pathogen seek healthcare, and a proportion that do seek healthcare can have a health event that is not captured within the system. Terminology for these events is often used in the literature with varying definitions (Gibbons et al., 2014). Here, we are interested in the degree of under-reporting by any cause and if a simple method can accurately adjust reported cases to represent the true number of infections.

To address the magnitude and extent of uncertainty within under-reporting, past researchers have used community-based studies (e.g., Sypsa et al., 2011), serological surveys (e.g., de Melker et al., 2006), returning traveler studies (e.g., Ekdahl & Giesecke, 2004), and capture-recapture studies (e.g., McCarty et al., 1993), all of which require some interaction with individuals and protected health information. Gibbons et al. (2014) presented an extensive review from 1990 to 2011 on all studies, using the methods previously mentioned, for two diseases (salmonellosis and campylobacteriosis). They extracted a multiplication factor, or the factor by which the reported number of disease cases must be multiplied to obtain an estimate of the actual number of cases. They found that methodology varied by study and geographic location leading to great uncertainty in the multiplication factor values that estimate the true incidence of each disease. While most literature studies may acknowledge under-reporting, the factor is typically not calculated, potentially biasing disease parameter estimates and resulting downstream calculations of important metrics (e.g., the basic disease reproduction number, R_0). Often, the calculation of the true multiplication factor for disease incidence under-reporting is near impossible in many situations due to the inability to perform the methods listed above. Simulations coupled with infectious disease models to explicitly estimate the fraction of under-reporting are an alternate route for addressing uncertainty. For example, Aronna et al. (2022) used modeling to estimate ~90% unreported cases for the first wave of the COVID-19 pandemic for Rio de Janeiro, Brazil.

Infectious disease models have many model classifications and categories with the original and most notable being the Susceptible – Infected – Removed (SIR) type models. SIR-based models continue to play an important role in mechanistically describing various types of epidemics (e.g., flu, COVID-19, sexually transmitted diseases, and others; see, e.g., Bagal et al. (2020); Postnikov (2020); Saad-Roy et al. (2020) and in producing predictions to inform policy decisions (Kermack & McKendrick, 1927; Edelstein – Keshet (1988); Murray, 1989). This approach is not without flaws for estimating under-reporting, but the simplicity of SIR models with very few parameters to estimate makes it an attractive choice. Often complex model parameters cannot be estimated reliably from the available data simply because the data are insufficient to accurately estimate large numbers of parameters, and the model is subject to overfitting. In addition, model parameters taken from literature sometimes are not characteristic of the location for which the estimates are to be applied Reed et al (2015); Aronna et al (2022).

Understanding how infectious disease data are collected and reported can also be a factor in cases involving under-reporting, as well as in model parameter estimation, bias and misclassification (Kalachev et al., 2022). Typically, reported data have the following case definition types: total infections, new infections, active infections, isolated or quarantined, recovered, and deaths. This terminology for collected and reported data types can be inconsistent across the literature and health systems. For example, in the well-known and frequently studied SIR-modeled flu outbreak in an English boarding school in 1978 (Edelstein – Keshet (1988); Murray, 1989), 763 school boys between the ages of 10 and 18 experienced the H1N1 flu. For modeling purposes, the boys were reported as "confined to bed", classically modeled as infected, and "convalescent", usually dropped from the analysis. However, Kalachev et al. (2022) showed how prediction accuracy and parameter estimation can be influenced by data misclassification, and argued that for the flu data, the sum of "confined to bed" and "convalescent", should be defined as isolated (or quarantined) cases. This highlights the fact that individual behavior can play an important role in spreading the disease (active spreaders vs. infected but isolated/self-isolated), and further complicate case definitions and model accuracy.

Here, we present a simple method which allows for estimating the fraction of reported cases for infectious diseases using the modified SIR model and a reporting parameter, k. We define k as the fraction of all infections reflected in reported cases. Then the under-reporting fraction f can be found by the formula $f = 1 - k$; this is the fraction of all infections missed somehow by a health system. Reported cases are assumed to be those individuals who were symptomatic, sought healthcare, and were correctly diagnosed (see the hypothetical example in Table 1). We illustrate the method with simulated data of known k values (i.e., that a certain percentage of cases are being reported). We then apply the method to estimate the fraction of reported in two empirical datasets for flu and COVID-19. We hypothesize that for common yearly flu the fraction of cases that are not reported is expected to be large since usually it is only those who end up in a hospital or become sick enough to see a doctor who are counted. On the other hand, one would expect that a vast majority of COVID-19 cases should have been reported in the fall of 2020, during the first wave of the disease in Missoula, given the heightened public awareness and lack of home testing kits at that time. The main question that we address in this paper may be stated as follows: if for a certain infectious disease only a fraction of the infections were reported, can this fraction be reliably estimated from the available "new cases" data (data most commonly reported for public records)?

Table 1
Illustrative example is given to represent those cases that are reported in health systems and those cases that are missed. Reported cases result from the proportions of infected who are symptomatic, who attend healthcare, and who are correctly diagnosed, and thus reported. In this example, 90% of all infected individuals become symptomatic, of whom 60% attend healthcare, of whom 55% are reported through the notification system. If the total (true) number of infections was 10,000, then the parameter k is 0.297 or 2970 reported number of infected individuals. It is important to emphasize that only the cumulative effect of the morbidity surveillance pyramid stages resulting in coefficient k can be estimated. The separate effect of individual stages cannot usually be elucidated with the available data.

	All Infections: 10,000 people			
	Symptomatic (90%): 9000 people			Asymptomatic (10%): 1000 people
Attending healthcare	YES (60%): 5400		NO (40%): 3600	NO: 1000
Correctly diagnosed	YES (55%): 2970	NO (45%): 2430	NO: 3600	NO: 1000
Reported	YES: 2970	NO: 2430	NO: 3600	NO: 1000
Fraction of all infections	$k = 0.9 \bullet 0.6 \bullet 0.55 = 0.297$	$f = 1 - k = 1 - 0.297 = 0.703$		
	Reported cases: 2970 cases (k can be estimated)	**Not reported (under-reported) cases: 2430 + 3600 + 1000 = 7030 cases (f = 1 - k can be estimated)**		

2. Materials and methods

2.1. Original SIR model, R_0, and minimal immunization requirement

We start with the well-known classical SIR model formulated for three (time dependent) populations, Susceptible, $S(t)$, Infected (active symptomatic and asymptomatic infection spreaders), $I(t)$, and Removed $R(t)$ (combining the recovered and those infected who are isolated/quarantined/self-isolated). We consider an infection spread in a community (region such as a city, county, state, etc.) with a constant total population, N, so that the conservation relationship holds, that is, births and deaths are equal or their difference is negligible (often the case for small and medium size communities over characteristic time comparable to one wave of an epidemic):

$$S(t) + I(t) + R(t) = N. \tag{1}$$

Then, the corresponding system of ordinary differential equations (ODEs) describing the behavior of $S(t)$ and $I(t)$ (due to (1), $R(t) = N - S(t) - I(t)$) can be written as follows:

$$\frac{dS}{dt} = -\alpha \bullet S \bullet I; \quad \frac{dI}{dt} = \alpha \bullet S \bullet I - \beta \bullet I; \tag{2}$$

with initial conditions:

$$S(0) = S^*, I(0) = I^*, \tag{3}$$

where S^* is the initial number of Susceptible, I^* is the initial number of Infected, and $R(0) = N - S(0) - I(0) = 0$, assuming the beginning of a disease outbreak and no prior vaccination. The proportionality constant α in (2) is related to the mobility of the population in the systems, i.e., to how many contacts each person on average can have per unit of time for a given population density at specified location, as well as to the infectious disease “effective” virulence, i.e., the likelihood of a new infection appearing per contact. The term $(-\alpha \bullet S \bullet I)$ in the first equation of (2) and a similar term $\alpha \bullet S \bullet I$ in the second equation of (2) describe the process of population transfer from the Susceptible compartment to the Infected (infection spreaders) compartment which, according to the Law of Mass Action, is proportional to the product of population densities in the respective compartments and which, in turn, can be re-written in the form of the actual populations' product. The term $(-\beta \bullet I)$ in (2) describes the rate of removal of individuals from the Infected compartment to the Removed compartment (containing individuals who are no longer able to spread the disease due to either isolation or recovery); β is the rate constant of this process, so that the characteristic “effective” removal time is proportional to $1/\beta$.

Equation (2) with conditions (3) do not have an exact explicit solution, but they have been extensively studied both qualitatively and quantitatively; see, e.g., (Edelstein – Keshet (1988); Murray, 1989). In principle, if sufficient data are available and defined correctly, the model parameters α and β can be easily estimated (Kalachev et al., 2022; Prodanov, 2021), which in turn, can be used to calculate the basic disease reproduction number, R_0, for a given disease:

$$R_0 = \frac{\alpha \bullet N}{\beta}. \tag{4}$$

The metric R_0 has been used as an indicator of infection spread (for R_0 greater than one, the infection will spread; for R_0 smaller than one, the infection dies out; see, (Edelstein – Keshet (1988); Murray, 1989). From here, the calculation can be made for the minimal fraction, p_0, of the initial population that would need to be immunized to avoid further infection spread

Edelstein – Keshet (1988); Murray, 1989; i.e., the fraction p of the population that has to be immunized must satisfy the inequality:

$$p > p_0 = 1 - \frac{1}{R_0} = 1 - \frac{\beta}{\alpha \bullet N}. \tag{5}$$

2.2. Modified SIR model with k, the proportion of reported cases

We define parameter k as the fraction or proportion of infections reflected in reported cases, or those individuals that were symptomatic, sought healthcare, and were correctly diagnosed and thus, reported (see the hypothetical example in Table 1). Next, we consider the situation where the disease propagation dynamics are adequately described by model (2) and initial conditions (3) with the conservation relationship (1), but in the case where a fraction of Removed cases was reported. Let us also assume that the fraction $0 \le k \le 1$ of actual reported cases, i.e., the reporting fraction, is a constant throughout a comparatively short period of an epidemic or single wave event. Then $f = 1-k$ (with $0 \le f \le 1$) will correspond to the fraction of cases not reported (or under-reported). This means that, instead of $R(t)$, the available data represent $k \cdot R(t)$, where k is unknown. In principle, we can either modify the data using k and fit the original SIR model to them, or we can modify the model and fit it to the original available data. Here, we present the latter.

We introduce the new re-scaled variables $\overline{S} = k \bullet S, \overline{I} = k \bullet I, \overline{R} = k \bullet R$. We assume that the proportion of reporting is the same among each of these groups in the population (e.g., the smaller reported number of Removed would be associated with the smaller reported number of Infected, etc.). This is guided by our goal of obtaining a modified SIR model with minimal changes that is mathematically equivalent to the original SIR model. The modified model is obtained as a result of equivalent transformations that can be reversed. Thus, the modified SIR is obtained from system (2) with corresponding initial conditions (3) and can be written as follows:

$$\frac{d\overline{S}}{dt} = -(\alpha / k) \bullet \overline{S} \bullet \overline{I}; \quad \frac{d\overline{I}}{dt} = (\alpha / k) \bullet \overline{S} \bullet \overline{I} - \beta \bullet \overline{I}; \tag{6}$$

$$\overline{S}(0) = k \bullet S^*, \overline{I}(0) = k \bullet I^* = \overline{I}^*, \tag{7}$$

and $\overline{R}(t) = k \bullet N - \overline{S}(t) - \overline{I}(t)$, which is equivalent to the original conservation relationship $R(t) = N - S - I$ multiplied by a constant k.

With regard to parameter estimation using nonlinear least squares implemented in numerous software packages (e.g., MATLAB, 2020), we note that compared to the original SIR model (1)–(3), the modified SIR model has an additional parameter k dividing the parameter α, and multiplying the initial Susceptible population value of S^*. The initial number of Susceptible S^* is assumed known (approximated); however, since we are under the assumption that the true number of infected is not known, the initial number of infected individuals, $\overline{I}^*$ must also be estimated during the model fitting process. In summary, the unknown parameters for estimation include $\overline{I}^*$, k, the re-scaled $\overline{\alpha} = \alpha/k$ and original (not re-scaled) β.

Justification of the proposed modified SIR model formulation is presented in the Appendix.

2.3. SIR model data types needed to estimate model parameters

Because of the importance of misclassification for modeling purposes, we note data definitions as follows. The most commonly available data are referred to in public records as “new cases”, and they represent the newly recorded disease cases which are reported at a given time unit (daily, weekly, etc.). We note that “new cases” cannot be interpreted simply as time dependent Infected since after being recorded, a person stays infected for a number of days. For some infectious diseases, e.g., COVID-19 in early pandemic waves, it is natural to interpret the time dependent sum of “new cases” as Removed since for this novel virus, as soon as a person is identified as sick and reported as a “new case”, they either self-isolate or are quarantined (depending on the severity of disease symptoms) and, thus, are removed from the active spreaders pool. In what follows, we consider situations where the “new cases” data are known (either simulated or obtained from the public domain) and interpreted as Removed for the purpose of classical SIR or modified SIR model fits (see Kalachev et al., 2022).

2.4. Creating simulated data for the true number of infections with various levels of k

To check how well the modified SIR model parameters can be estimated in the presence of under-reporting, we produced a simulated data set with the following parameter values: $\alpha = 0.000015$ [1/(person • week)], $\beta = 1.00$ [1/week], $I^* = 100$ [persons], and $N = 100,000$ people. The numerical solution of the SIR model with exact parameter values was constructed. Random errors (1% normally distributed) were then added to the newly weekly Removed counts. These error-induced weekly counts were then added for consecutive weeks to generate the cumulative time dependent Removed cases. As a result, the errors for the time dependent Removed portion of the population were guaranteed to satisfy the following natural condition:

for every consecutive week the generated cumulative number of Removed was greater than or equal to the generated cumulative number of Removed obtained for the previous week. Simulated data sets were generated in the manner described above for three different fractions k of reporting ($k = 0.01, 0.5, 0.99$); the resulting non-integer population numbers were rounded to the nearest integer; the modified SIR model was fit to the noisy simulated data for the three choices of reporting parameter. We note that the size of the errors added to the exact simulated data to make them more realistic was chosen to be small (1%) so that the modified SIR model fits could capture a very low (1%) and a very high (99%) fraction of reported cases. The simulated data were produced using the MATLAB software package (MATLAB, 2020) and are included in the Supplementary Materials (**Data A.1**).

2.5. Fitting the empirical datasets to the modified SIR model

The 2016–2017 flu data for Missoula County, Montana, USA, courtesy of Montana Department of Public Health and Human Services, cover the period from October 24, 2016 to June 4, 2017 and correspond to 32 full weeks (**Data A.2**). The COVID-19 data for Missoula County, Montana, USA, courtesy of Missoula City-County Health Department, cover August 17, 2020–December 20, 2020 and correspond to 18 full weeks (**Data A.3**; Kalachev et al., 2022). For the flu data, the number of new cases was reported and for the COVID-19 data, the numbers of new and active cases were reported. For both examples, only the new cases data were used to estimate the fraction of reported cases. Where data included daily counts, we reduced noise by using weekly data to fit models. Removed cases starting from week one were obtained by adding the new cases observed on a given week to the previously observed new cases (the sum of new cases for all the previous weeks). The choices of empirical datasets are representative examples for a single pandemic wave cycle (in the case of COVID-19) and a single epidemic seasonal cycle (in the case of flu). For other (pre-COVID-19) flu seasons the data fitting results are similar, with small variations in estimated model parameter values due to differences in flu strains, and slight differences in patterns of human behavior that may change from one year to another.

The Missoula County populations in 2016 and 2020 were estimated to be 115,983 and 121,630 people, respectively (https:townfolio.co). According to information from Montana Department of Health and Human Services, the vaccination coverage in Montana for flu was at the level of 42.2% (https:dphhs.mt.gov), slightly lower than national averages. We assume that this percentage was also indicative of vaccination coverage for Missoula County. According to US Centers for Disease Control and Prevention (CDC) the effectiveness of the flu vaccine during the 2016–2017 flu season was 40% (flu/about). So, based on these vaccination numbers, the initial Susceptible population in Missoula County at the beginning of the 2016–2017 flu season was estimated to be $N = 115,983 - (0.422) \bullet (0.4) \bullet (115,983) = 96,405$. For the first wave of COVID-19, we assume no immune population exists and the number of deaths was negligible compared with the total County population.

For the flu epidemic and for the COVID-19 pandemic cases, the modified SIR models with reporting parameter k were fitted to under-reported cumulative "new cases" treated as under-reported Removed: $\overline{R}(t_j)$, where t_j identify weeks when the data were recorded/reported. In addition to k, the model parameters α, β, and I^* were also estimated. All computations included in this paper (direct solutions of the SIR model to produce the simulated data sets, SIR model fits to data using appropriate optimization programs, etc.) were made using the MATLAB software package (MATLAB, 2020) and corresponding toolboxes.

3. Results

3.1. Simulation study of reporting fraction k using modified SIR model

The resulting estimates of modified SIR model parameters and corresponding standard errors for the simulated data cases are provided in Tables 2–4. We note that the true values of the parameters in these tables are those which enter the original SIR model used for simulations. They could have been estimated by simply fitting the original SIR model to the data if 100% of the infections were reported. In the case of under-reporting these parameter values were estimated using the modified SIR model introduced above.

Let us emphasize that the estimated model parameter values are close to those for which the simulated data were produced for all three choices of reporting coefficient k. Unsurprisingly, the greatest disagreement between the true and estimated values occurs when $k = 0.01$. We would not expect them to be exactly the same because artificial variability (1% relative error), as described earlier, was introduced in the simulated data to replicate noise in real data. In addition to Estimated Values (EV) and Standard Errors (SE), the ratios EV/SE for each parameter are presented to indicate the magnitudes of the SEs relative to the estimated parameters.

Table 2
Simulated data: estimated parameter values and standard errors for the case of $k = 0.01$.

Parameter	Units	Estimated Value (EV)	Standard Error (SE)	EV/SE
k	non-dim.	0.010296	0.000224	45.9
α	1/(person·week)	1.547e-05	3.690e-07	41.9
β	1/week	1.0491	0.0373	28.1
I^*	persons	93.888	4.4734	20.9

Table 3
Simulated data: estimated parameter values and standard errors for the case of $k = 0.5$.

Parameter	Units	Estimated Value (EV)	Standard Error (SE)	EV/SE
k	non-dim.	0.50024	7.2489e-05	6900.9
α	1/(person·week)	1.5007e-05	2.6632e-09	5635.1
β	1/week	1.0008	0.0002	5003
I^*	persons	99.899	0.0145	6895.2

Table 4
Simulated data: estimated parameter values and standard errors for the case of $k = 0.99$.

Parameter	Units	Estimated Value (EV)	Standard Error (SE)	EV/SE
k	non-dim.	0.99006	0.000122	8087.6
α	1/(person·week)	1.5001e-05	2.2711e-09	6605
β	1/week	1.0001	0.00017	5860.2
I^*	persons	99.994	0.0124	8085.5

3.2. Estimating k for the 2016–2017 Missoula County flu data

Here, we present the results of the Modified SIR model fit to weekly reported new flu cases from Missoula County, MT, for the 2016–2017 flu season. The results are shown in Figs. 1 and 2, and Table 5.

In Fig. 1, the cumulative reported "new cases" data interpreted as Removed cases are presented together with the estimated re-scaled Infected (active infection spreaders) $\bar{I}(t_j)$, estimated re-scaled Removed $\bar{R}(t_j)$, and re-scaled total cases $(k \bullet N - \bar{S}(t_j))$. In Fig. 2 we show the predictions for the actual numbers of Infected $I(t)$, Removed $R(t)$, and total cases $N - S(t)$. Note that these predictions are roughly 100 times the fitted numbers of cases shown in Fig. 1, indicating that only about 1% of flu cases were actually reported.

As with the simulation, we emphasize that Table 5 contains the estimates of the original model parameter values rather than the re-scaled values. From the results presented in Table 5, as supported by Fig. 2, the model estimates indicate that only about 1.2% of flu cases were reported; i.e., 98.8% of cases were estimated to not have been reported. For the estimated parameter values, parameter α is the rate constant of the process describing the appearance of newly infected per unit characteristic time (here, per week) and per already infected individual placed in the pool of current Susceptible. The characteristic time associated with the appearance of newly infected, $1/(\alpha \cdot N)$, is estimated to be 0.7336 of a week, or about 5.1 days. Parameter β is the rate constant of the process describing the removal of individuals from the Infected pool (due to isolation, self-isolation and recovery); the characteristic "effective" removal time, $1/\beta$, is estimated to be 1.0706 weeks, or about 7.5 days, which is an approximate flu recovery time. The estimated initial number of Infected I_0 during the "zero" week, approximately equal to 99, represents the "average" number of active infection spreaders observed each day during the first week of the epidemic.

For estimated disease metrics, the basic disease reproduction number, given by (4) is estimated to be $R_0 = 1.4594$. The corresponding minimal immunization requirement (i.e., the fraction of people who have to be immunized to avoid an epidemic), given by (5), is estimated to be $p_0 = 0.31,$ or 31% of the vulnerable population (under the condition that the vaccine is 100% effective).

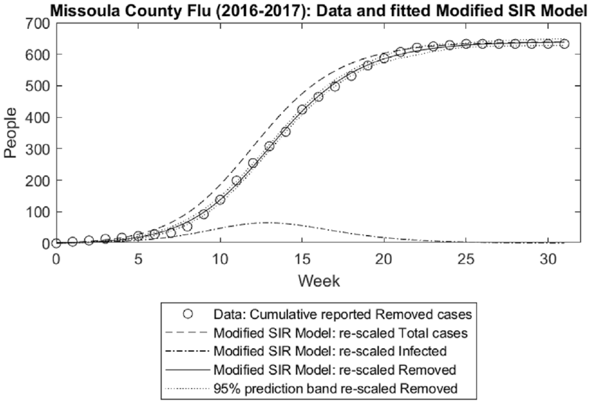


Fig. 1. Weekly flu data for Missoula County, MT; modified SIR model fit.

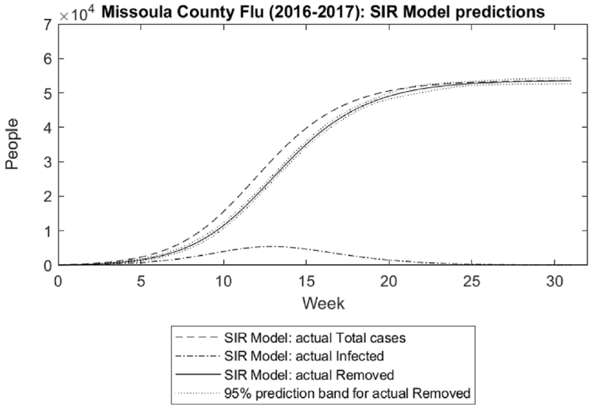


Fig. 2. SIR Model predictions for 2016–2017 flu data in Missoula County, MT.

3.3. *Estimating k for Missoula County COVID-19 data*

Next, we apply the modified SIR model to COVID-19 “new cases” data reported for Missoula County in the fall of 2020. In Fig. 3 we show the results of the modified SIR model fit to the available cumulative “new cases” data interpreted as Removed. Model parameter estimates with corresponding standard errors are shown in Table 6.

Once again, the weeks in the figure are counted from August 17, 2020, which roughly coincides with the onset of the Delta variant in Missoula County. Note that with these data, the re-scaled Total cases and re-scaled Removed counts are virtually identical due to the high isolation/quarantine rate during the Covid-19 pandemic. As with the flu model, Table 6 displays the original model parameter estimates, i.e., not the re-scaled estimates.

The results presented in Table 6 indicate that slightly more than 99% of COVID-19 cases were reported, and taking into account the estimated standard error for the reporting fraction k we may state that the k-value obtained by the model fit to these data is not statistically different from 1 (100% reporting of cases).

The basic disease reproduction number is estimated to be $R_0 = 1.0288,$ with corresponding minimal immunization requirement of $p_0 = 0.028$ or 2.8%. We also note that the characteristic time associated with the appearance of newly infected, $1/(\alpha\cdot N)$, is estimated to be 0.0904 of a week, or about 0.633 of a day, which is much shorter compared to that estimated for the flu during the 2016–2017 flu season. The characteristic “effective” removal time, $1/\beta$, is estimated to be 0.0879 weeks, or about 0.6153 of a day, which is also much shorter compared to that observed during the flu season of 2016–2017.

For comparison, let us fit the original SIR model, containing just three parameters (α, β, I^*; and no k, which is equivalent to setting $k = 1$ in the modified SIR model) to the available data. Model parameter estimates with corresponding standard errors are shown in Table 7.

Not surprisingly, in the case of the original SIR model, the estimates of α, β, and I^* closely agree with those obtained for the modified SIR model, likely because k was estimated to be so close to 1. The fitted Removed cases curve in the case of the original SIR model produces the same results as those shown in Fig. 8. Corresponding estimates for R_0 and p_0 are close to (8) and (9), respectively. The Mean Squared Error for the original SIR model fit is 1958.67 and for the modified SIR Model it is 2089.22. The small-sample Akaike's Information Criterion (AICc) indicates that the original SIR model is better (more informative) compared to the modified one in describing the same COVID-19 data set. In particular, for the original SIR model we have AICc $=$ 194.0612$+C$, which has a delta AICc more than 2 points smaller than that for the modified one, for which AICc $=$ 197.4236$+C$; here C is a constant (same for both models).

Therefore, the simpler original SIR model containing just three parameters appears to be better than the modified SIR model for representing the collected data in this case and for making predictions on propagation of the COVID-19 infection under conditions similar to those observed in Missoula County in the fall of 2020.

4. Discussion and conclusion

4.1. *A modified SIR model produces a simple solution for estimating the reporting fraction*

A simple method for estimating the fraction of reported cases during infection spread was introduced and applied to real data obtained for Missoula County in 2016–2017 (flu season) and in the fall of 2020 (COVID-19 pandemic). The method is based on using a modified version of the classical SIR model under the assumption that model parameters do not change drastically during one wave of a particular epidemic/pandemic. This assumption should hold for small and mid-size communities (population <200,000) which exhibit stable patterns of population behavior over the duration of an epidemic/

Table 5
Modified SIR model fit results: estimated parameter values and standard errors (2016–2017 flu in Missoula, MT, USA).

Parameter	Units	Estimated Value (EV)	Standard Error (SE)	EV/SE
k	non-dim.	0.011927	0.000226	52.8
α	1/(person·week)	1.4139e-05	6.4145e-07	22.0
β	1/week	0.9340	0.0009	1012.7
I^*	persons	98.58	11.10	8.8

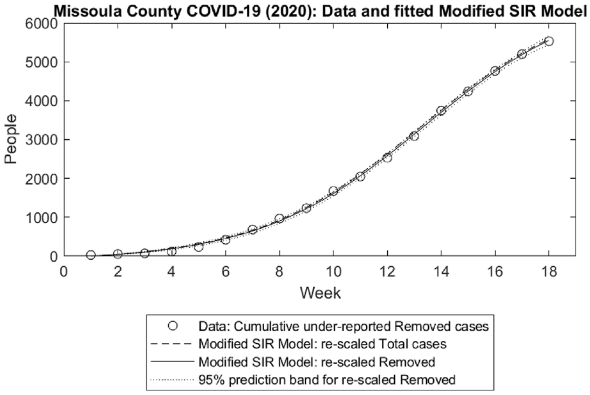


Fig. 3. Weekly COVID-19 data for Missoula County, MT; modified SIR model fit.

Table 6
Modified SIR model fit results: estimated parameter values and standard errors (COVID-19 in fall of 2020 in Missoula County, MT, USA).

Parameter	Units	Estimated Value (EV)	Standard Error (SE)	EV/SE
k	non-dim.	0.992	0.0471	21.1
α	1/(person·week)	9.3538e-05	8.8516e-06	10.6
β	1/week	11.059	0.039	283.6
I^*	persons	3.5004	0.2336	15.0

Table 7
Original SIR model fit results: estimated parameter values and standard errors (COVID-19 in fall of 2020 in Missoula County, MT, USA).

Parameter	Units	Estimated Value (EV)	Standard Error (SE)	EV/SE
k	non-dim.	1	0	–
α	1/(person·week)	9.4173e-05	3.0174e-07	312.1
β	1/week	11.1364	0.0378	294.6
I^*	persons	3.4261	0.0551	62.2

pandemic, as was illustrated for Missoula County, Montana, USA data. For larger communities (e.g., an entire country) the assumption of stable patterns of population behavior requires further investigation.

The proposed approach produced precise estimates of the reporting parameter k for the simulated data sets and for both discussed real data sets. It is important to comment further on the practical meaning of k. In general, the process resulting in under-reporting of infection cases involves several stages, including only a fraction of cases being symptomatic, another fraction seeking medical care, yet another fraction being correctly diagnosed, etc. (Table 1). The reporting fraction k (and related fraction of cases not reported, $f = 1 - k$) , which we estimate in this paper, may be considered as a combination (a product) of a number of such fractions characteristic of various stages of a health care reporting chain. To estimate each individual fraction mentioned above, additional data must be collected and reported.

This proposed method for estimating the level of reporting/under-reporting can currently only be applied to single waves of infections, e.g., a single wave cycle in the case of COVID-19 or a single seasonal cycle in the case of flu. The seasonal reporting varies from year to year due to many factors including the characteristics of circulating viruses, the timing of the season, population immunity to circulating viruses, how well flu vaccines are working, and how many people have gotten vaccinated (https://www.cdc.gov/flu). Therefore, the estimate of k by this proposed method would need to be recalculated for each wave. In addition, age groups have been shown to vary in their reporting Reed et al (2015) whereas our method estimates reporting for the entire population. Below we present a discussion of the particular results and conclusions following from our analysis.

4.2. Under-reporting for a case where k nears 0

During the 2016–2017 flu season in Missoula County, we estimate that ~1.2% of the cases were reported. For Missoula County this corresponds to an overall burden of influenza of an estimated 53,572 illnesses. The United States Centers for Diseases Control and Prevention (CDC) estimates the burden of influenza in the United States from a multiplier method described in Reed et al. (2015) and (Rolfes et al., 2018) retrospectively to 2010. Their estimates of illness burden used an adjustment of hospitalization rates which uses the percent of persons hospitalized with respiratory illness who were tested for influenza and the average sensitivity of influenza tests used in the participating FluSurv-NET hospitals. Based off of these data, the overall burden of influenza for the 2016–2017 season was an estimated 29 million influenza illnesses, 14 million influenza-associated medical visits, 500,000 influenza-related hospitalizations, and 38,000 influenza-associated deaths (https://www.cdc.gov/flu). On the other hand, the total number of positive flu tests reported by various laboratories for the 2016–2017 flu season was 43,232 (https://www.cdc.gov/flu/weekly). This means that an estimated $43{,}232/29{,}000{,}000 \times 100\% \approx 0.15\%$ of the total estimated flu cases were actually reported to CDC. Our analysis shows comparable, but slightly higher, reported numbers with many possible explanations, including different (not directly comparable) data sources, and different methods of analysis of reporting/under-reporting estimates, etc. We note that our estimate is also close to that obtained for early stages of 2009 H1N1 flu pandemic in Mexico using a combination of surveillance data, statistics of travelers arriving from Mexico to different counties and modeling approach: the fraction of reported cases estimated in (Chong et al., 2014) was about 0.46%. In (McCarthy et al., 2020) a dynamic disease transmission model, laboratory-confirmed influenza surveillance data, and randomized-controlled trial (RCT) data were used to quantify the underestimation factor for flu epidemics in the USA and Canada during the 2011–2012 and 2012–2013 flu seasons. Estimates of under-reporting generated from modified CDC-defined flu (influenza) in RCT include: 0.5% (USA, 2011–2012); 0.27% (USA, 2012–2013); 2.6% (Canada, 2011–2012); and 1.2% (Canada, 2012–2013).

The CDC estimates that influenza vaccination during the 2016–2017 influenza season prevented 5.3 million illnesses, 2.7 million medical visits, 72,000 hospitalizations and 5200 deaths associated with influenza (https://www.cdc.gov/flu/vaccines-work). Our modified SIR model indicated that the flu epidemic could have been avoided with a certain level of additional vaccination. We estimated that $p_0 = 0.31,$ or 31% of the vulnerable population would need to be additionally vaccinated. We use the term 'additionally vaccinated' since we already adjusted for 42.2% vaccination rates along with 40% vaccine efficacy, or $115{,}983 \cdot 0.422 \cdot 0.40 = 19{,}578$ individuals were assumed to be immune and protected from the flu, producing 96,405 individuals that were assumed to be vulnerable and used as the initial Susceptible population for modeling. Therefore, the additional vaccinations needed would be 31% of the assumed vulnerable population and accounting for 40% vaccine efficacy, this is $96{,}405 \cdot 0.31/0.4 = 74{,}714$ individuals. However, with the original 42.2% vaccinated (48,944) and the additional 74,714 needed, we see that this exceeds 100% of the Missoula County population and the flu epidemic could not have been avoided. It is important to point out that the total Missoula County population is an estimate and therefore the numbers presented above are approximate to help illustrate the overall effect of under-reporting on vaccination rates needed to stop flu propagation dynamics within a particular flu season.

4.3. Under-reporting for a case where k nears 1

During the first wave of COVID-19 in Missoula County, we estimate that ~99% of the cases were reported, a stark contrast with the 2016–2017 flu season. This higher percentage is likely caused by a heightened public awareness around a novel and emergent virus. For example, during the fall of 2020 Missoula County had strict social distancing and mask wearing measures along with thorough case testing, tracing, and isolation. This analysis indicates that during fall of 2020, if a 100% effective vaccine would have been available, the epidemic could have been stopped by immunizing just about 3% of the county population under these conditions (if Missoula County were to be isolated from the rest of the country, i.e., no travel to and from the county, etc.).

Other studies have estimated under-reporting for COVID-19 as well. With a similar multiplier method as described above for influenza, the CDC estimates approximately 25% of cases were reported from the beginning of the pandemic in the US through September 2021 (https://www.cdc.gov/coronavirus). However, this estimate spans many waves across and through time in the US, as our analysis spanned only the first wave within Montana, USA. A serologic survey of SARS-CoV-2 antibodies in 10 geographically diverse US sites from 23 March to 12 May of 2020 estimated that the total number of SARS-CoV-2 infections reported was ~10% (Havers et al., 2020). Aronna et al. (2020) also estimated ~10% of Brazil's cases were reported for a similar time frame (March–July 2020). Our high estimate of reported cases over other literature could be due to many reasons, including a comparatively small Missoula County population size and thus, relatively easier capabilities to conduct a more thorough case tracing, testing, and isolation.

4.4. Use of simple versus modified SIR model and caveats of this study

This study illustrates two extreme case studies for under-reporting where $k = 0.012$ and $k = 0.99$ for flu and COVID-19, respectively. Because the fraction of reported cases in the COVID-19 data was close to 1, to describe the infection spread during the COVID-19 pandemic in the fall of 2020, it was sufficient to use the classical SIR model rather than a modified SIR model. Regardless, the use of a simple versus modified SIR model represents a population in three categories (2–3 parameters) and does not fully account for other disease transition rates (asymptomatic, death, exposed, etc.) or age-structured populations. However, due to the lack of data available to fit those other compartments, we argue for Occam's Razor in selecting simple models to assess.

For methodological purposes, let us address in more detail the proposed model's limitations, connection to other model formulations, and potential model extensions. Over the years, the original SIR model proved to be a reliable tool for describing the propagation of various infectious diseases and for making predictions on the expected number of affected individuals (for making healthcare policy decisions), on minimal vaccination requirements leading to elimination of epidemics, just to name a few (Edelstein – Keshet (1988); Murray, 1989). The advantage of using SIR models is their simplicity. However, this simplicity could be misleading because, depending on the availability of data and on the situation being modeled, the interpretation of model parameters and understanding the meaning of modeling results may become non-trivial. Below we present an explanation of several controversial items which are most often encountered in connection with the use of SIR models.

(a) *One of the major concerns about using the classical SIR models is related to seemingly not being able to take into account the incubation period of an infection, and not including explicitly a corresponding compartment* (e.g., *for Exposed*) *in the original or modified SIR model formulation.* In the classical SIR models the parameter α describes the combination of characteristic times associated with both human behavior (i.e., the average number of human interactions over some characteristic time interval, at a given location being studied) and the virulence of an infection, which is related to the probability of an individual being infected per encounter and the incubation period of a disease. At first glance, it seems to be very logical to split the combined process of getting infected into two components, one responsible for human behavior, and the other responsible for properties of a virus. This will naturally lead to the so-called SEIR (Susceptible-Exposed-Infected-Removed) model formulation. However, there is a practical problem with such an extension. There are no data available which would allow one to untangle the two processes and reliably estimate two parameters instead of just one, α. The ad hoc heuristic assumptions on the incubation period time, especially for a new virus, may lead to completely erroneous prediction results due to extreme model sensitivity to such arbitrary assumptions (the characteristic time associated with the incubation period must be estimated as a part of the model fitting process, but the necessary corresponding data are usually unavailable). The estimated “combined” parameter α can still provide us with important information: in particular, if the virus remains the same (e.g., a flu during a given season, or a certain COVID-19 strain during a particular wave of a disease) and the estimated parameter values for α turn out to be approximately the same for several locations (which is the case, e.g., for different cities in Montana), this would mean that the patterns of human behavior at these locations during epidemics are similar and, consequently, could be taken into account for modeling of future epidemics caused by different new types of viruses. The epidemics in different locations (e.g., within the state of Montana) do not happen simultaneously. Sometimes a delay between the start of the infection waves (for the same virus) in two Montana cities could be as long as six weeks. This means that the model parameters obtained for the city where an epidemic started earlier may be used to reliably predict the propagation of infection in a city where an epidemic is going to start with a delay. This approach to using the results of data driven SIR modeling has direct practical implications for the health care related policy decisions which now can be made based on the actual data collected for an actual epidemic.

(b) *Another concern frequently expressed is related to so-called asymptomatic infection cases and ways these could be included in various types of infection propagation models.* The presence of potentially asymptomatic infection spreaders is certainly a problem for the classical SIR model (which requires the 100% of cases to be reported and data to be fitted for estimating model parameters) since the asymptomatic are not counted in the reported cases, but still spread the infection. This issue is naturally handled by the modified SIR model proposed in this paper since asymptomatic cases may just be treated as a portion of the under-reported cases. Indeed, for modeling purposes, there is no difference between a symptomatic individual who was identified by the system but refused to be isolated/quarantined, and an asymptomatic individual, who does not realize they are sick and continues roaming among the Susceptible spreading the infection. In the modified SIR model discussed in this paper the meaning of the parameter β is tied to the characteristic time (given by $1/\beta$) of removal of individuals from the active infection spreaders pool. The paths of such removal may vary (e.g., a person may simply recover from a disease and stop transmitting an infection as a result, or a person may be quarantined, and then stop spreading an infection while at the same time remaining infected, etc.) and the effective value of the parameter β will reflect some effective characteristic time $1/\beta$ averaged for all the removal paths involved. Thus, if no isolation/quarantine is imposed on the infected population, $1/\beta$ will approximately correspond to the characteristic duration of a disease; if isolation/quarantine of infected individuals is mandated, then $1/\beta$ will approximately correspond to the characteristic time between appearance of the first symptoms of a disease and imposed isolation. Naturally, the numerical value of $1/\beta$ must belong to an interval between 0 days (which corresponds to immediate isolation) and average duration of a disease (7–14 days for flu/COVID-19). If the asymptomatic cases are present, the estimated value of β will reflect this fact, and the more asymptomatic cases there are, the closer the characteristic time $1/\beta$ will turn out to be to an average individual's disease duration. The estimated value of parameter β therefore may be used as a quality control for deciding how well the SIR model fits the available data: if this value is unrealistically small, i.e., $1/\beta$ turns out to be unrealistically large (like a month or a year), this will indicate that something is wrong either with the data, with the model interpretation, or with both.

(c) *One additional concern involves the use of disease data for comparatively small communities, like Missoula, Montana, and the potential for extension of the results to larger cities, countries, etc.* Currently about 33.5% of the US population, or about 112 million people (https://data.oecd.org), live in so-called small urban areas (with population between 50,000 and 200,000 residents). The presented analysis was intentionally performed for one such area to prove the concept and illustrate how the analysis works under "ideal" conditions. For such small communities it is easier to collect reliable data, to implement various prevention measures, to check the consistency of patterns of population behavior, and to separate both temporally and spatially different waves of a disease. This is often not possible to do for large cities and countries considered as a whole, where more complex patterns of human behavior during epidemics may be observed, where different waves of a disease could be lumped together, where spatially separate events are often reflected in the same data set, etc. We took advantage of highly reliable flu and COVID-19 data sets collected for Missoula, where one of the important conditions for the proposed model to work appears to be satisfied: i.e., the pattern of human interactions did not change much during one wave of a disease (flu or COVID-19). This condition limits the application of the proposed approach to a particular kind of disease, i.e., to respiratory infections with comparatively short duration (that is why when several waves of a disease corresponding, e.g., to different variants of a virus are observed, a separate analysis must be performed for each wave). The presented analysis is expected to perform very well for the small urban areas mentioned above, which is very promising because the reliable predictions obtained by the proposed modified SIR model may, in principle, affect health policy decisions for millions of people in small urban areas. For larger cities additional studies must be performed on the applicability of the proposed algorithm.

We conclude by stating that the presented approach of estimating the reporting parameter k can be easily applied to the extended versions of the modified SIR model which, in principle, may include an Exposed compartment, age structured populations, population travel between various locations, etc., as long as the corresponding data on Exposed, on new cases reported for various age groups, on travel records, etc., become available. The presented approach is purely data driven. The model parameter estimates are performed using only available data for a given location under minimal and very mild assumptions.

Ethics approval

No ethical approval is required.

Funding

This work was supported by National Institute of General Medical Sciences of the National Institutes of Health, United States (Award numbers P20GM130418 and U54GM104944).

CRediT authorship contribution statement

Leonid Kalachev: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Jon Graham:** Writing – review & editing, Writing – original draft, Investigation, Funding acquisition, Conceptualization. **Erin L. Landguth:** Writing – review & editing, Writing – original draft, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

Authors declare that they have no competing interests.

Acknowledgments

The authors thank Missoula City County Health Department for collecting the data used in this study and their dedication and commitment in responding to the COVID-19 pandemic. We also would like to thank anonymous reviewers for their very helpful comments and suggestions which led to a considerable improvement of the final version of the manuscript.

Appendix

Justification of the proposed rescaled SIR model formulation.

Here we discuss in more detail the mathematical aspects of modeling under-reporting of cases, specifically for the situations when SIR-type models are used for the analysis. We explain our approach for estimating the reporting fraction and justify our choice of SIR model re-scaling (i.e., the modified SIR model) presented in the previous subsections.

We emphasize once again that one condition must be satisfied for the proposed approach to work properly: the rate of under-reporting must be consistent, i.e., it must be (approximately) the same, over the time period corresponding to one wave of an infection/epidemic/pandemic. This condition is expected to hold for a given location for a comparatively short period (from one to several weeks, corresponding to one wave of the disease).

To correctly align the available data, often called “new cases”, with the appropriate compartment of the SIR model, let us re-write this model in the following (equivalent) form involving differential equations for Infected (active symptomatic and asymptomatic infection spreaders), $I(t)$, and Removed $R(t)$ (combining the recovered and those infected who are isolated/quarantined/self-isolated). First, we note that Susceptible, $S(t)$, are expressed using a constant total population, N, and the population conservation relationship (1):

$S(t) = N - I(t) - R(t)$.

Then, the corresponding system of ordinary differential equations (ODEs), equivalent to (2), describing the behavior of $I(t)$ and $R(t)$ can be written as follows:

$$\frac{dI}{dt} = \alpha \bullet (N - I - R) \bullet I - \beta \bullet I; \quad \frac{dR}{dt} = \beta \bullet I. \tag{8}$$

The initial conditions are

$$I(0) = I^*, R(0) = 0, \tag{9}$$

where I^* is the initial number of Infected, and $R(0) = 0$, assuming the beginning of a disease outbreak and no prior vaccination or immunity; then, the initial number of Susceptible is $S^* = N - I^*$. The notation and meaning of the parameters are the same as those introduced in the previous sections.

To better understand the structure of the phase trajectories of system (8) on the (R,I)-phase plane let us rescale the time variable using the characteristic removal time $1/\beta$: the new non-dimensional time will be $\tau = t \bullet \beta$. Thus, the new rescaled system (8) depends only on one parameter, α/β:

$$\frac{dI}{d\tau} = \left(\frac{\alpha}{\beta}\right) \bullet (N - I - R) \bullet I - I; \quad \frac{dR}{d\tau} = I. \tag{10}$$

Initial conditions (9) hold for (10) as well. The corresponding phase trajectories (uniquely defined for the given initial conditions and for the value of α/β characteristic of a given location and given disease) on the (R,I)-phase plane are described by the differential equation, obtained by dividing the first equation of (10) by the second equation, and with initial condition following from (9):

$$\frac{dI}{dR} = \left(\frac{\alpha}{\beta}\right) \bullet (N - I - R) - 1; \quad I(R = 0) = I^*. \tag{11}$$

We note that the shapes of the phase trajectories do not change if the original parameters α and β change proportionately, i.e., if instead of α and β we substitute into (10) the values $\rho\alpha$ and $\rho\beta$, where ρ is some constant, then $(\rho\alpha)/(\rho\beta) = \alpha/\beta$, and the form of (10) stays the same! For a given location and disease type, i.e., for fixed values of α and β, different phase trajectories are chosen by specifying different initial conditions, i.e., choices of $I(t = 0) = I^*$ and $R(t = 0) = R^*$, so that $I(R = R^*) = I^*$.

In addition to the model description, we now need to discuss the meaning of the available data, usually referred to in the media and research literature as “new cases”. For the majority of the infectious diseases associated with viruses, like flu, the data are collected opportunistically, and only the patients with severe infection cases seeking medical attention, and consecutively placed in a hospital/isolated/quarantined are recorded as new cases. For the common flu, according to the CDC, the reporting fraction of seasonal flu cases is usually below 1% (for example, for the 2016–2017 flu season it was estimated to be 0.15%). For COVID-19, the first symptoms of disease immediately led to isolation/self-isolation/quarantine of the patients (i.e., reporting was expected to be close to 100%). So, for both diseases discussed in our paper, the "new cases" may safely be interpreted as the rate at which the Infected were converted to Removed and, thus, eliminated from the active infection spreaders pool (if the "new cases" data are under-reported then the rate of conversion is under-reported as well). We note that for other infectious diseases this may not necessarily be the case; a separate discussion is needed to verify the correct interpretation of "new cases" data and the alignment of these data with the appropriate model compartment(s). In the presence of under-reporting, there are at least four equivalent ways to fit the actually recorded "new cases" data using an SIR model (all of them only work under an assumption of an approximately constant under-reporting fraction over the duration of a disease wave): one can rescale the available data using the originally unknown reporting/under-reporting parameter, and then (a) fit the Removed from the SIR model to rescaled cumulative "new cases" or (b) fit the rate of change of Removed to rescaled "new cases"; or one may instead rescale the variables and parameters in the SIR model (which is preferable since the data are usually considered to be "given", and independent of any model, but the models may be naturally modified by scientists studying disease propagation) and (c) fit the rescaled Removed from the modified SIR model to the available cumulative "new cases" data or (d) fit the rate of change of rescaled Removed to the available "new cases" data. In all four cases the reporting parameter is determined as a part of the model fitting procedure. In our paper, approach (c) is chosen (see the body of the text), which is much easier to explain mathematically (see below), and which explicitly uses one of the SIR model's compartments, Removed, for data fitting. Let us explain how the available data may be visualized in the phase plane of system (10) in more detail.

For some particular choices of α and β three sample phase trajectories of system (10), whose shapes are described by (11) with various initial conditions, are shown in Fig. 4.

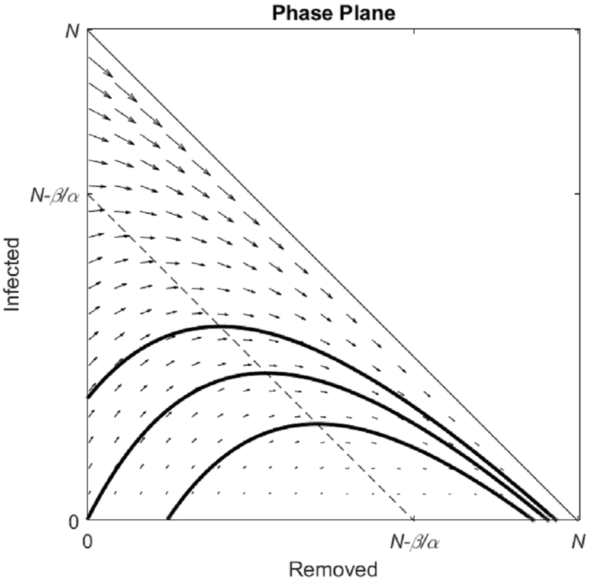


Fig. 4. Sample phase trajectories of system (10). Sample velocity vectors illustrate the vector field associated with (10); the solutions of (10) correspond to the points in the phase plane moving along the trajectories in the directions specified by the velocity vectors. We note that the phase trajectories are located in the first quadrant of the (R, I)-coordinate system, and they belong to a triangular region bounded by the horizontal and vertical axes and a segment of the straight-line N=$I(t)+R(t)$. A dashed line $N-I-R=\beta/\alpha$ separates the above mentioned region into two domains: in one of them $\frac{dI}{d\tau}>0$, and in the other $\frac{dI}{d\tau}<0$.

Each disease wave for a given location produces a unique phase trajectory. Assume that all the parameters in the model are known, together with the initial conditions corresponding to a certain epidemic. Then exactly one phase trajectory describes the SIR model solution behavior for this epidemic. Also, assume that 100% of "new cases" are reported. Then, to visualize these data, consider time dependent cumulative "new cases" obtained by adding up recorded daily/weekly "new cases" for consecutive days/weeks (the equally spaced non-dimensional time points τ_i at which the data were collected are easily obtained from the original dimensional time points t_i via simple rescaling: $\tau_i = t_i \bullet \beta$). The corresponding data points are shown in Fig. 5; they are just the projections of the "exact" solution points $(R(\tau_i), I(\tau_i))$, belonging to the phase trajectory, onto the Removed axis or, in other words, the R-coordinates of the "exact" solution points at $\tau = \tau_i$ on the (R,I)-phase plane. These points are spaced according to the equal time intervals between data recordings (we note that equal time intervals correspond, in general, to unequal "spatial" intervals traveled by the solution point along the phase trajectory during these equal time intervals).

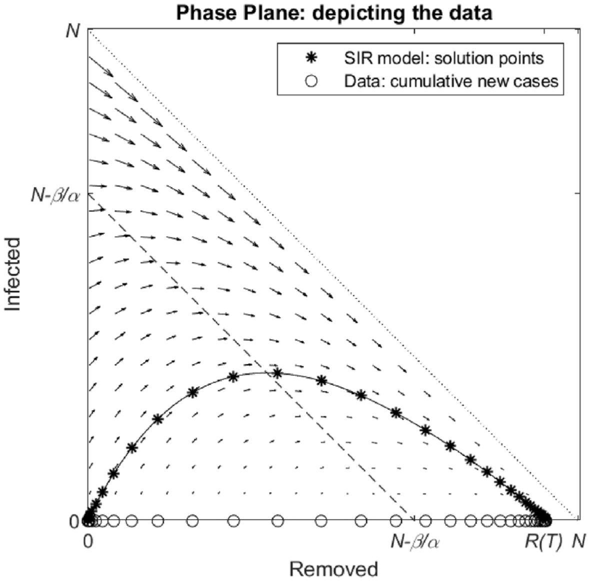


Fig. 5. One phase trajectory of system (10) obtained for a particular choice of initial conditions. The simulated cumulative "new cases" data (circles) corresponds to projections of the simulated solution points data (stars) onto the Removed axis (R-axis).

Sample exact 100% reported "new cases" and cumulative "new cases" data, corresponding to the situation illustrated in Fig. 5 depicting the (R,I)-phase plane, are shown in Fig. 6(a), and the consistently under-reported "new cases" and cumulative "new cases" data are shown in Fig. 6(b); see further discussion presented below. The latter is also presented in the phase plane illustrated in Figs. 7 and 8.

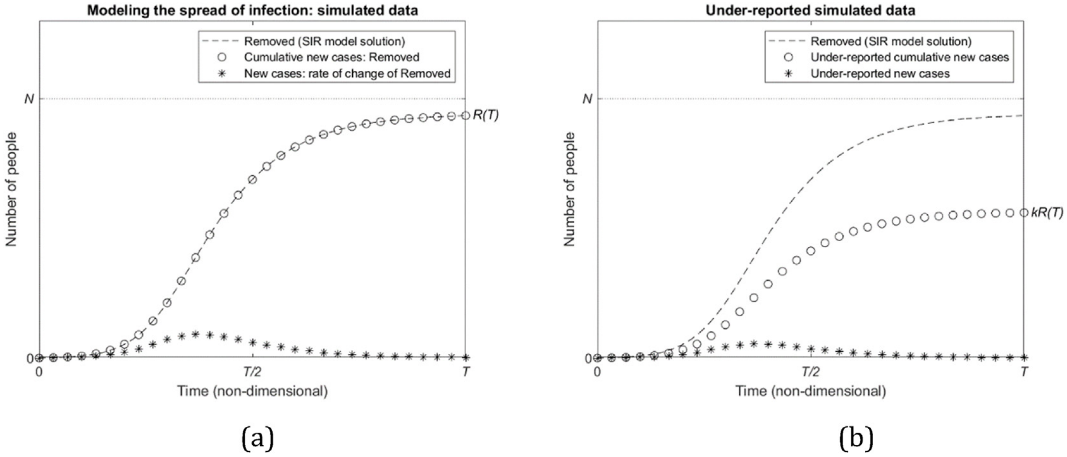


Fig. 6. (a) Simulated data for the case of 100% reporting. (b) Simulated data for the case of under-reporting. Here, Removed data are indicated by circles. $R(T)$ is the total number of Removed / Recovered at the end of the epidemic; k is the infection reporting fraction; T is the time corresponding to the end of the epidemic, so that $I(T) = 0$. See additional explanation in the text.

Let us mention that the shape of the phase curve does not depend on the direction of the time variable. That is, if the time is reversed in system (10), i.e., the independent variable $-\tau$ is substituted for τ, the differential equation (11) will not change. Since the initial number of infected is rarely known exactly and it is often estimated as a part of the data fitting procedure, it is natural to construct the phase trajectories using the end of the epidemic as a starting point for solving (11). Indeed, at the end of an epidemic, i.e., at $t = T$, it is known that $I(T) = 0$ and $R(T)$ is the total number of removed (or eventually recovered), so that the initial condition for (11) is $I(R(T)) = 0$; this number, $R(T)$, becomes $k \bullet R(T)$ in the presence of under-reporting, where k is the reporting fraction.

Thus, to construct the phase trajectory corresponding to the SIR model solution with under-reported number of cases one must use the following "initial" condition corresponding to the known "end-of-the-epidemic" numbers: $I(k \bullet R(T)) = 0$, and solve (11) backwards with respect to R, for $R < k \bullet R(T)$, and stop integration at the value of $R = R^*$ for which $I(R^*) = 0$; naturally $R^* \in [0, k \bullet R(T)]$. It can be easily seen that if the cumulative "new cases" data are under-reported, i.e., $k \bullet R(\tau_i)$ are provided instead of $R(\tau_i)$, no phase trajectory satisfying such data exists that could be produced by the original (not modified) SIR model. Indeed, let us choose some value of the reporting coefficient $0 < k < 1$. Then the phase trajectory describing the solution of the original SIR model corresponding to the under-reported data solving (11), constructed for $R < k \bullet R(T)$, and satisfying the condition $I(k \bullet R(T)) = 0$ only exists for the case where $N - \frac{\beta}{\alpha} < k \bullet R(T) < R(T)$; a sample trajectory of this type is shown in Fig. 7; it obviously misses a large number of available data points corresponding to the beginning of an epidemic. A phase trajectory satisfying (11) and the above mentioned initial condition does not exist in the first quadrant in the case where $0 < k \bullet R(T) \le N - \frac{\beta}{\alpha}$; this case is illustrated in Fig. 8. As a result, either rescaling of the data, or the rescaling of the original SIR model is required to correctly fit the data and estimate model parameters. In this paper we chose the model rescaling (reversible transformation of model parameters and variables leading to an equivalent model formulation which we call the *modified SIR model*) in the form shown in the body of the text.

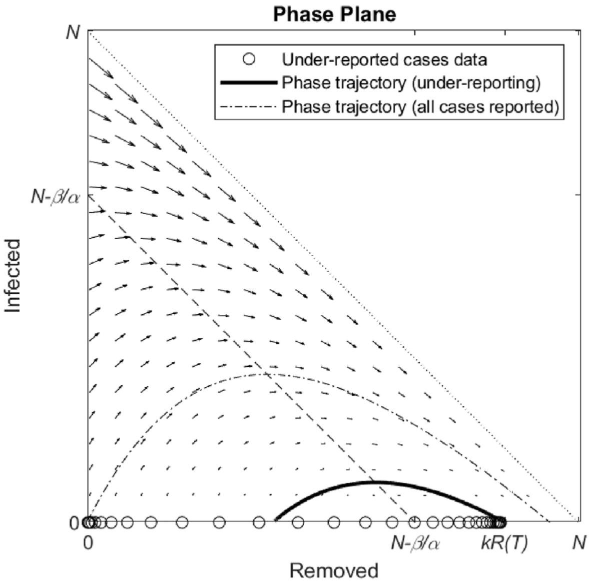


Fig. 7. Phase trajectory of the original (not modified) SIR model in the case of under-reported data (solid curve), which satisfy the condition $N - \frac{\beta}{\alpha} < k \bullet R(T) < R(T)$. Here T is the time corresponding to the end of the epidemic.

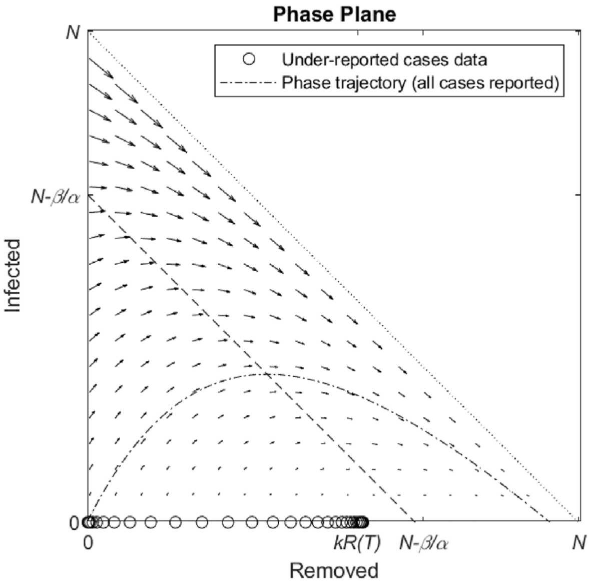


Fig. 8. Illustration of the fact that no phase trajectory of the original (not modified) SIR model exists in the case of under-reported data, which satisfy the condition $0 < k \bullet R(T) \le N - \frac{\beta}{\alpha}$. Here T is the time corresponding to the end of the epidemic.

Let us reiterate that for a particular location and a given disease the shapes of the phase trajectories produced by the corresponding SIR model are completely determined by the combination of parameters α/β, the actual phase trajectory corresponding to a particular wave of infection as chosen by the initial conditions (the initial number of Infected I^*). The parameter β determines the spacing of the data recordings collected at certain (in our case, equally spaced) times. If 100% of disease cases are reported, the cumulative "new cases" data associated with the Removed compartment of the SIR model allows one to reliably estimate three parameters: $\alpha, \beta,$ and I^*. If the data are consistently under-reported, the modified SIR model, containing one additional parameter, the fraction of reported cases k, allows one to reliably estimate four parameters: α, β, I^*, and k.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idm.2024.06.002.

References

Aronna, M. S., Guglielmi, R., & Moschena, L. M. (2022). Estimate of the rate of unreported COVID-19 cases during the first outbreak in Rio de Janeiro. *Infectious Disease Modelling, 7*, 317–332.

Bagal, D. K., Rath, A., Barua, A., & Patnaik, D. (2020). Estimating the parameters of susceptible-infected-recovered model of COVID-19 cases in India during lockdown periods. Chaos. *Solitons and Fractals, 1400*, Article 110154.

Chong, K. C., Fong, H. F., & Zee, C. Y. (2014). Estimating the incidence reporting rates of new influenza pandemics at an early stage using travel data from the source country. *Epidemiology and Infection, 142*, 955–963.

https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html.

https://data.oecd.org/popregion/urban-population-by-city-size.htm.

de Melker, H. E., Versteegh, F. G. A., Schellekens, J. F. P., Teunis, P. F. M., & Kretzschmar, M. (2006). The incidence of Bordetella pertussis infections estimated in the population from a combination of serological surveys. *Journal of Infection, 53*(2), 106–113. https://doi.org/10.1016/j.jinf.2005.10.020

https://dphhs.mt.gov/assets/publichealth/CDEpi/Infographics/InfluenzaCoverageinMontana20162017.pdf.

Edelstein – Keshet, L. (1988). *Mathematical models in Biology*. New York: Random House.

Ekdahl, K., & Giesecke, J. (2004). Travellers returning to Sweden as sentinels for comparative disease incidence in other European countries, campylobacter and giardia infection as examples. *Euro Surveillance, 9*(9), 6–9.

https://www.cdc.gov/flu/about/burden/2016-2017.html.

https://www.cdc.gov/flu/vaccines-work/burden-averted-2016-17.htm.

https://www.cdc.gov/flu/weekly/weeklyarchives2016-2017/Week39.htm.

Gibbons, C. L., Mangen, M. J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., & Kretzschmar, M. E. E. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: A comparison of methods. *BMC Public Health, 14*, 147. https://doi.org/10.1186/1471-2458-14-147

Havers, F. P., Reed, C., Lim, T., Montgomery, J. M., Klena, J. D., Hall, A. J., , … Owen, S. M., et al. (2020). Seroprevalence of antibodies to SARS-CoV-2 in 10 sites in the United States, March 23-may 12, 2020. *JAMA Internal Medicine, 180*(12), 1576–1586. https://doi.org/10.1001/jamainternmed.2020.4130

Kalachev, L., Graham, J., & Landguth, E. L. (2022). *Classical infectious disease modelling paradigms challenged by the SARS-CoV-2 pandemic, accepted for publication in Infectious Disease Modeling*.

Keramarou, M., & Evans, M. R. (2012). Completeness of infectious disease notification in the United Kingdom: A systematic review. *Journal of Infection, 64*, 555–564.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society. Series A, 115*, 700–721.

McCarthy, Z., Athar, S., Alavinejad, M., Chow, C., Moyles, I., Nah, K., , … Taurel, A.-F., et al. (2020). Quantifying the annual incidence and underestimation of seasonal influenza: A modelling approach. *Theoretical Biology and Medical Modelling, 17*(11). https://doi.org/10.1186/s12976-020-00129-4

McCarty, D. J., Tull, E. S., Moy, C. S., Kwoh, C. K., & LaPorte, R. E. (1993). Ascertainment corrected rates: Applications of capture-recapture methods. *International Journal of Epidemiology, 22*(3), 559–565. https://doi.org/10.1093/ije/22.3.559

Murray, J. D. (1989). *Mathematical Biology*. Berlin: Springer-Verlag.

Postnikov, E. B. (2020). Estimation of COVID-19 dynamics "on a back-of-envelope": Does the simplest SIR model provide quantitative parameters and predictions? *Chaos, Solitons & Fractals, 135*, Article 109841.

Prodanov, D. (2021). Analytical parameter estimation of the SIR epidemic model. Applications to the COVID-19 pandemic. *Entropy, 23*, 59.

Reed, C., Chaves, S. S., Daily Kirley, K. P., Emerson, R., Aragon, D., Hancock, E. B., Butler, L., Baumbach, J., et al. (2015). Estimating influenza disease burden from population-based surveillance data in the United States. *PLoS One, 10*(3), Article e0118369.

Rolfes, M. A., Foppa, I. M., Garg, S., Flannery, B., Brammer, L., Singleton, J. A., … Reed, C. (2018). Annual estimates of the burden of seasonal influenza in the United States: A tool for strengthening influenza surveillance and preparedness. *Influenza Other Respir Viruses, 12*(1), 132–137. https://doi.org/10.1111/irv.12486. Epub 2018 Feb 14.

Saad-Roy, C. M., Wagner, C. E., Baker, R. E., Morris, S. E., Farrar, J., Graham, A. L., Levin, S. A., Mina, M. J., Metcalf, C. J. E., & Grenfell, B. T. (2020). Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years. *Science, 370*(6518), 811–818. https://doi.org/10.1126/science.abd7343. Epub 2020 Sep. 21.

Sypsa, V., Bonovas, S., Tsiodras, S., Baka, A., Efstathiou, P., Malliori, M., Panagiotopoulos, T., Nikolakopoulos, I., & Hatzakis, A. (2011). Estimating the disease burden of 2009 pandemic influenza A(H1N1) from surveillance and household surveys in Greece. *PLoS One, 6*(6), Article e20593. https://doi.org/10.1371/journal.pone.0020593

Tam, C. C., Rodrigues, L. C., Viviani, L., Dodds, J. P., Evans, M. R., Hunter, P. R., Gray, J. J., Letley, L. H., Rait, G., Tompkins, D. S., & O'Brian, S. J. (2012). Longitudinal study of infectious intestinal disease in the UK (IID2 study): Incidence in the community and presenting to general practice. *Gut, 61*(1), 69–77. https://doi.org/10.1136/gut.2011.238386

https://townfolio.co/mt/missoula-county/demographics.

https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm.

Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Jr., Reingold, A., Arnold, B. F., Hubbard, A., & Benjamin-Chung, J. (2020). Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications, 11*, 4507.