PLOS ONE

RESEARCH ARTICLE

Cure and death play a role in understanding dynamics for COVID-19: Data-driven competing risk compartmental models, with and without vaccination

Min Lu, Hemant Ishwaran¤*

Department of Public Health Sciences, Miller School of Medicine, University of Miami, Miami, FL, United States of America

¤ Current address: Department of Public Health Sciences, Miller School of Medicine, University of Miami, Don Soffer Clinical Research Center, Miami, FL, United States of America
* hemant.ishwaran@gmail.com

Citation: Lu M, Ishwaran H (2021) Cure and death play a role in understanding dynamics for COVID-19: Data-driven competing risk compartmental models, with and without vaccination. PLoS ONE 16(7): e0254397. https://doi.org/10.1371/journal.pone.0254397

Editor: Yury E Khudyakov, Centers for Disease Control and Prevention, UNITED STATES

Received: February 2, 2021

Accepted: June 25, 2021

Published: July 15, 2021

Data Availability Statement: We used only public data for this analysis, which was obtained from New York Times COVID-19 data for the U.S., 2020 available online at: https://github.com/nytimes/covid-19-data This work does not include individual participant data and does not involve the use of identifiable health information.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Several factors have played a strong role in influencing the dynamics of COVID-19 in the U.S. One being the economy, where a tug of war has existed between lockdown measures to control disease versus loosening of restrictions to address economic hardship. A more recent effect has been availability of vaccines and the mass vaccination efforts of 2021. In order to address the challenges in analyzing this complex process, we developed a competing risk compartmental model framework with and without vaccination compartment. This framework separates instantaneous risk of removal for an infectious case into competing risks of cure and death, and when vaccinations are present, the vaccinated individual can also achieve immunity before infection. Computations are performed using a simple discrete time algorithm that utilizes a data driven contact rate. Using population level pre-vaccination data, we are able to identify and characterize three wave patterns in the U.S. Estimated mortality rates for second and third waves are 1.7%, which is a notable decrease from 8.5% of a first wave observed at onset of disease. This analysis reveals the importance cure time has on infectious duration and disease transmission. Using vaccination data from 2021, we find a fourth wave, however the effect of this wave is suppressed due to vaccine effectiveness. Parameters playing a crucial role in this modeling were a lower cure time and a signficantly lower mortality rate for the vaccinated.

Introduction

The coronavirus disease (COVID-19) first identified in Wuhan, China in late 2019 [1] rapidly spread across mainland China, and then across the globe, eventually manifesting itself into the global pandemic of 2020. Evidence from regions affected early in the pandemic suggested a high fatality rate, which combined with the highly infectious nature of COVID-19, spurred both national and international governments to impose strict measures to reduce its spread.

These early attempts to contain the outbreak were partially effective. For example, lockdown measures implemented in the U.S. in March and early April of 2020 was undoubtedly instrumental in reducing number of infected cases [2]. However, due to pressure to relieve economic hardship and revive economies, imposed sanctions began to be lifted [3]. In the U.S., beginning in mid April of 2020, a number of states began lifting lockdown measures. This was followed by a dramatic increase in reported cases in many regions [4], which was then countered by returning to earlier stricter social distancing. After that a pattern of closing and opening up state economies ensued in reaction to perceived waves. Then in late December of 2020, COVID-19 vaccines became available to some of the U.S. population and a mass vaccination program began in earnest in 2021.

These factors have contributed to a complex dynamic process and because of this, many things remain unclear. For example, estimated fatality rates of early COVID-19 data have been reported to be 6–20% [5], but more recent data suggests far lower numbers. For the U.S., what is the rate? Has it changed, and if so when did this occur? Another matter are waves. Wave-like behavior of COVID-19 has been reported as matter of fact, but how many waves have occurred, when did they occur and what are their characteristics? A related issue is the metric used to measure fatality. We note that values 6–20% reported above are from calculations by David Baud and colleagues [5], which they referred to as a "mortality rate", but which are technically the cumulative death rate, defined as the proportion of a group dying within a specified time interval [6]. A more widely adopted measure is case fatality rate, defined as the proportion of deaths due to a specific disease over total number of diseased cases relative to length of time. This work estimates a related value, the case fatality risk, abbreviated here as CFR, and defined as the probability of death for an infectious case. Here the term risk is used in place of rate, as rate refers to a specific time period, whereas risk refers to the probability of an adverse outcome [7]. The CFR is easily understood: *for an individual with COVID-19, what is the probability they will die?*

A widely used tool to study dynamics of an infectious disease are predictive epidemiological models [8–16]. One of the most commonly used of these is the SIR compartmental model [17]. This characterizes individuals of a population in terms of three stages of infection: susceptible, infected and recovered. Extensions to the basic SIR model to include other stages have been considered for COVID-19 [18–22]. An implicit assumption of compartmental models is an exponentially distributed infection time. Under this framework, most of the infected are assumed to recover or die early in the infection duration, which may not conform to observed COVID-19 survival behavior [23–25]. To overcome this, extensions to SIR rely on multiple stages using mathematical models with additional parameters; however these can be difficult to estimate especially with limited data available in epidemic scenarios. There has been some work to more directly attack this issue by using non-exponential distributions [26]. Lofgren et al. [22] grouped patients into several stages based on type of exposure and patient risk for analysis of fecal microbiota transplantation data. The integro-differential equation formulation [27, 28] and the method of stages [29–31] have been used for measles. However, these require the distribution to be nonincreasing [27] or assume the mortality in the exposed and infectious classes is ignorable [28, 30, 31].

While there has been a substantial effort to extend compartmental models for more realistic analysis of infectious data, an overlooked approach is directly addressing the competing forces at play that "remove" an infectious individual from a susceptible population. Current models do not make a distinction for removal, but when an individual becomes infected with COVID-19, they are removed due to one of two conditions occurring: death or cure. Thus at any given time, there is an instantaneous risk of either dying or being cured for the infected individual.

In survival analysis, this type of data is called competing risk data and there is a large literature that has been developed for addressing this.

We extend the classical epidemiological compartmental model by incorporating competing risks using flexible hazard models for cure and death that can be used with epidemiological data. This yields a continuous removal rate that is a function of time and allows for flexible modeling of the dynamic process. It also makes it possible to estimate survival parameters for the disease, such as the probability of dying from COVID-19 (the CFR).

As motivation, consider Fig 1 which displays summary statistics for the U.S beginning from January 21th 2020, through to February 1st 2021. This will be referred to as pre-vaccination data as most of this data is prior to the large scale vaccination efforts of 2021. The figure highlights two waves, one being the time period of May 9 through August 27, which as will be explained is a period believed to characterize the second wave of the epidemic (the first wave being onset of the disease in early February). During this second wave period, there were 4,592,625 confirmed cases and 103,349 deaths from COVID-19 compared with 1,291,641 confirmed cases and 77,380 deaths recorded as of May 8th [32]. A unique pattern characterizing this second wave is a higher disease incidence rate combined with a lower apparent case fatality ratio (aCFR). The aCFR is defined as cumulative deaths due to the infection divided by cumulative confirmed cases and is a useful statistical quantity as it approximates the CFR. Fig 1 also highlights a third wave that follows completion of the second wave. Following a decrease in incidence rate for the second wave, it is characterized by an increasing incidence rate relative to daily deaths.

Using the competing risk compartmental model, we analyze pre-vaccination data using a four-parameter lognormal model in combination with a data driven contact rate. To tune parameters we make use of empirical data, a technique that has proven to be effective for describing the complex dynamics of COVID-19 [33]. Our analysis confirms the presence of the previously described waves and also chararacterizes their properties. We find that second and third waves have a significantly decreased CFR 1.7% compared with 8.5% of the first wave and also that the third wave has a longer period and a higher contact rate than previous waves. A what-if analysis which studies how dynamics change with parameters, reveals the

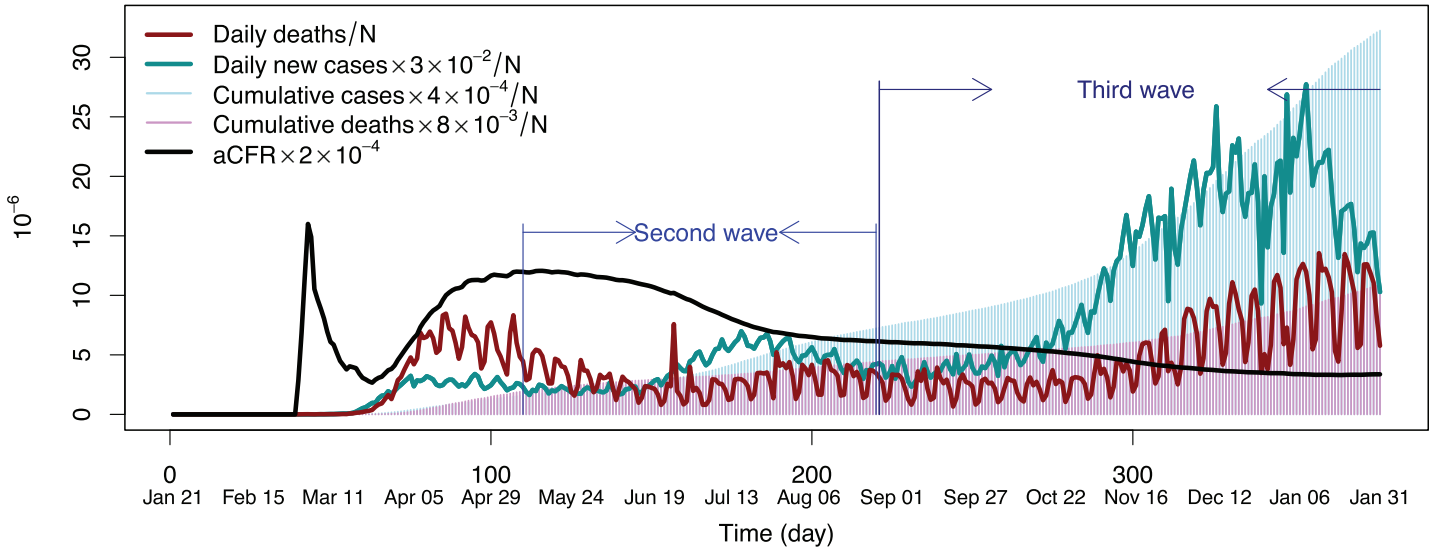

Fig 1. Statistics for COVID-19 pre-vaccination data in the U.S from January 21th, 2020 to February 1st, 2021. Note that values have been scaled in order to allow comparisons in one figure.

https://doi.org/10.1371/journal.pone.0254397.g001

importance of infectious time. This is crucial to understanding the effectiveness of vaccines. Extending the competing risk compartment model to include a vaccination component, we analyze 2021 vaccination data and find a fourth wave, however the effect of this wave is suppressed due to effectiveness of the vaccine. Parameters that play a crucial role in accurate modeling of this data are a lower cure time and significantly lower mortality rate for those vaccinated.

Materials and methods

Competing risk compartmental model

Let S denote number of susceptible, I the number of infected and R the number removed. We propose the following generalization of the SIR model [10]:

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta(t)IS}{N}\\ \frac{dI}{dt} &= \frac{\beta(t)IS}{N} - \gamma(t)I\\ \frac{dR}{dt} &= \gamma(t)I.\end{aligned}$$

Here $N = S + I + R$ is the total population which is assumed constant; thus the above set of equations reduces to two equations. The above generalizes the classical SIR model by allowing contact rate and removal rate to change with time. The value $\beta(t)$ is the contact rate at time t, equal to average number of contacts per person per time multiplied by probability of disease contact between a susceptible and infectious case at time t. The function $\gamma(t)$ denotes the removal rate at time t.

The removal rate is described as a function of t by utilizing a competing risk framework. We call this the Susceptible Infectious Cure Death (SICD) model. Let X be the continuous event time of an infected individual who either recovers from infection (is cured) or dies due to infection. The distributions for the two competing risks of X are specified using cause-specific hazard functions (one family that will be especially useful are lognormal distributed variables). Letting $h_C(x)$ and $h_D(x)$ denote the cause-specific hazards for cure and death (i.e. instantaneous risk of cure and death), we have the following key identity for the number removed (Theorem 1 in S1 Appendix),

$$\frac{dR}{dt} = \gamma(t)I = \gamma_C(t)I + \gamma_D(t)I. \tag{1}$$

The removal rate $\gamma(t)$ equals $\gamma_C(t) + \gamma_D(t)$ where γ_C and γ_D are the averaged h_C and h_D cause-specific hazards, averaged over length of time an infectious individual is infectious prior to t. This shows that the number of removed can be separated into number of cured and number dead in terms of the underlying hazards, which is an important feature exploited by our algorithm.

Identity (1) clarifies how the the choice of hazard effects the model. Consider the classical SIR model, which assumes X is exponentially distributed. Then $\gamma(t) = \gamma$ is a constant function and the cause-specific hazards for cure and death are also constant functions, $h_C(t) = \lambda_C$, $h_D(t) = \lambda_D$ (Corollary 1 in S1 Appendix). Let M_{death} be the the limit of the cumulative incidence function for death, where the latter is defined as the probability of an infectious individual experiencing death by a specified time (Definition 1 in S1 Appendix). Then M_{death} equals the CFR and for the exponential model we have $M_{\text{death}} = \lambda_D/\gamma$. Denoting the mean infectious period by $\bar{X}$, by the mean property of an exponential random variable, we have $\bar{X} = 1/\gamma$.

Therefore, $\lambda_D = M_{\text{death}}/\bar{X}$ and $\lambda_C = (1 - M_{\text{death}})/\bar{X}$, which shows that just fitting $(M_{\text{death}}, \bar{X})$ already uses up the two available degrees of freedom (λ_C, λ_D) for the model. This is one way to see why the classical model will be too inflexible for COVID-19 data.

Discrete time model

The SICD model is numerically implemented using a discrete time algorithm that takes both time t and infectious duration x as discrete intervals so that the solution can be calculated iteratively (see Section S1.3 in S1 Appendix). Days d are used in place of x for infectious duration time. To indicate discrete time for β a subscript of t is used. Values c_d and m_d are discrete versions for the cure and death hazards h_C and h_D.

The number of infectious cases $I(t)$ on day t is $I(t) = \sum_{d=0}^{\infty} i(t, d)$ where $i(t, 0)$ is the number of newly infected cases and $i(t, d)$ is the number of infectious cases on day t who have been infected for $d \geq 1$ days. The basic identity is

$$N = I(t) + S(t) + R(t) = I(t) + S(t) + [C(t) + D(t)]$$

where $R(t) = C(t) + D(t)$ is the total number removed and $C(t)$ and $D(t)$ are the total of all cured and dead up to day t.

Moving from day $t - 1$ to day t, the $I(t - 1)$ cases transmit disease to the susceptible cases at a discrete contact rate of β_t. Consequently, the number of newly infected cases on day t is

$$i(t, 0) = \beta_t \frac{I(t-1)S(t-1)}{N}.$$

There are three possible outcomes for the infectious cases on day $t - 1$ going to day t: cured, death, or infectious (status quo), with probabilities depending on infectious duration (Fig 2). For $i(t - 1, d)$, the probability of cure is c_d, the probability of death is m_d, and the probability of remaining infectious is $1 - c_d - m_d$. The infectious cases, $i(t - 1, d) \times (1 - c_d - m_d)$, will be counted as $i(t, d + 1)$ on day t because their infectious duration increases one day, i.e. $i(t, d + 1) = i(t - 1, d)(1 - c_d - m_d)$. The cured cases, $i(t - 1, d)c_d$, and deaths, $i(t - 1, d)m_d$, are counted towards daily cured and deaths on day t, yielding

$$\frac{dR}{dt} \asymp R(t) - R(t-1) = \sum_{d=0}^{\infty} i(t-1, d)c_d + \sum_{d=0}^{\infty} i(t-1, d)m_d.$$

Hence, solutions for $I(t) = \sum_{d=0}^{\infty} i(t, d)$, $R(t)$, and $S(t) = N - I(t) - R(t)$ can be obtained once we are given values $\{\beta_t\}_1^{T_{\max}}$, $\{i(0, d)\}_0^M$, and $(\{c_d\}_0^M, \{m_d\}_0^M)$; the latter are conditional cure and death rates obtained from the discretized hazards for h_C and h_D. Here M is a large number such that $i(t, d)$ can be assumed to be zero for $d > M$; thus sums are constrained to M terms. The value $T_{\max}$ equals maximum number of days under study.

Pre-vaccination model parameters and identification and observability

The discrete time algorithm was applied to New York Times COVID-19 data for the U.S. from January 21, 2020 to February 1, 2021 [32]. Population size was $N = 325{,}217{,}163$ equal to the sum of populations for the states and regions reported by the New York Times. Values were intialized using $i(0, 0) = 1$, and $D(0) = C(0) = i(0, 1) = i(0, 2) = \cdots = i(0, M) = 0$. A value of $T_{\max} = M = 377$ was used for the time window. Mean infectious duration was set to $\bar{X} = 29$ days: 14 or more days to develop symptoms [34], 7 days of moving average of the interval from symptom onset to isolation in hospital or quarantine [35], and 8 days from hospital admission to mortality or discharge (the average of 7 days for mortality and 9 for discharge) [23].

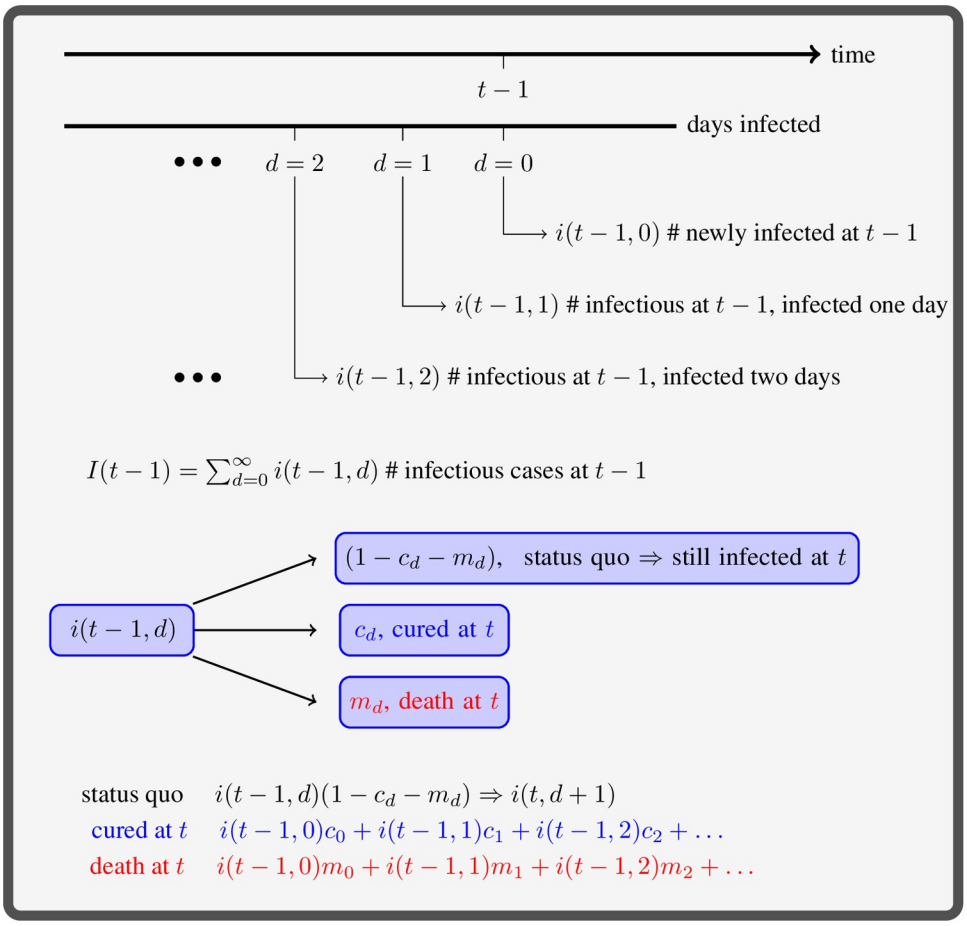

Fig 2. Infectious cases at time $t-1$ who have been infected d days, $i(t-1,d)$, have three possible outcomes at time t. An infectious case is either cured, they die, or they remain infectious, with rates c_d, m_d, and $1-c_d-m_d$, respectively.

https://doi.org/10.1371/journal.pone.0254397.g002

For exponential hazards, corresponding to the classical SIR model, parameters were set to $\lambda_C = (1-M_{\text{death}})/\bar{X}$ and $\lambda_D = M_{\text{death}}/\bar{X}$ (for the first wave in Fig 1 with 8.5% CFR this is $\lambda_C = 0.032$ and $\lambda_D = 2.93\times10^{-3}$). Lognormal cause-specific hazards were set to parameter values (μ_C, σ_C) for cure and (μ_D, σ_D) for death, where $\mu_C = 3.506$, $\sigma_C = 0.51$, $\mu_D = 3.8$, and $\sigma_D = 0.91$ (Section S2 in S1 Appendix). Regarding the issue of identifiability and observability, the classical SIR model is structurally identifiable with observable state $S(t)$ when either $I(t)$ or cummulative incidence data is used for the directly measured state [36]. These results continue to hold if the removal rate is a continuous time-varying function (see Model 6 from the S1 Appendix of [37]). Thus the SICD is identifiable with observable state $S(t)$. Later we will investigate the issue of practical identifiability for the SICD model.

Data-driven time varying contact rate

Data driven values were used for the discrete time contact rates β_t and set using the following approach. Let I_t^{new} denote the observed number of newly infected cases on day t. Because cases reported before $\bar{X}$ days are typically either cured or dead, the discrete contact rate β_t was estimated by the following: $\beta_t = I_t^{\text{new}} / \sum_{(s=t-\bar{X})}^{t} I_s^{\text{new}}$ (Fig 3A).

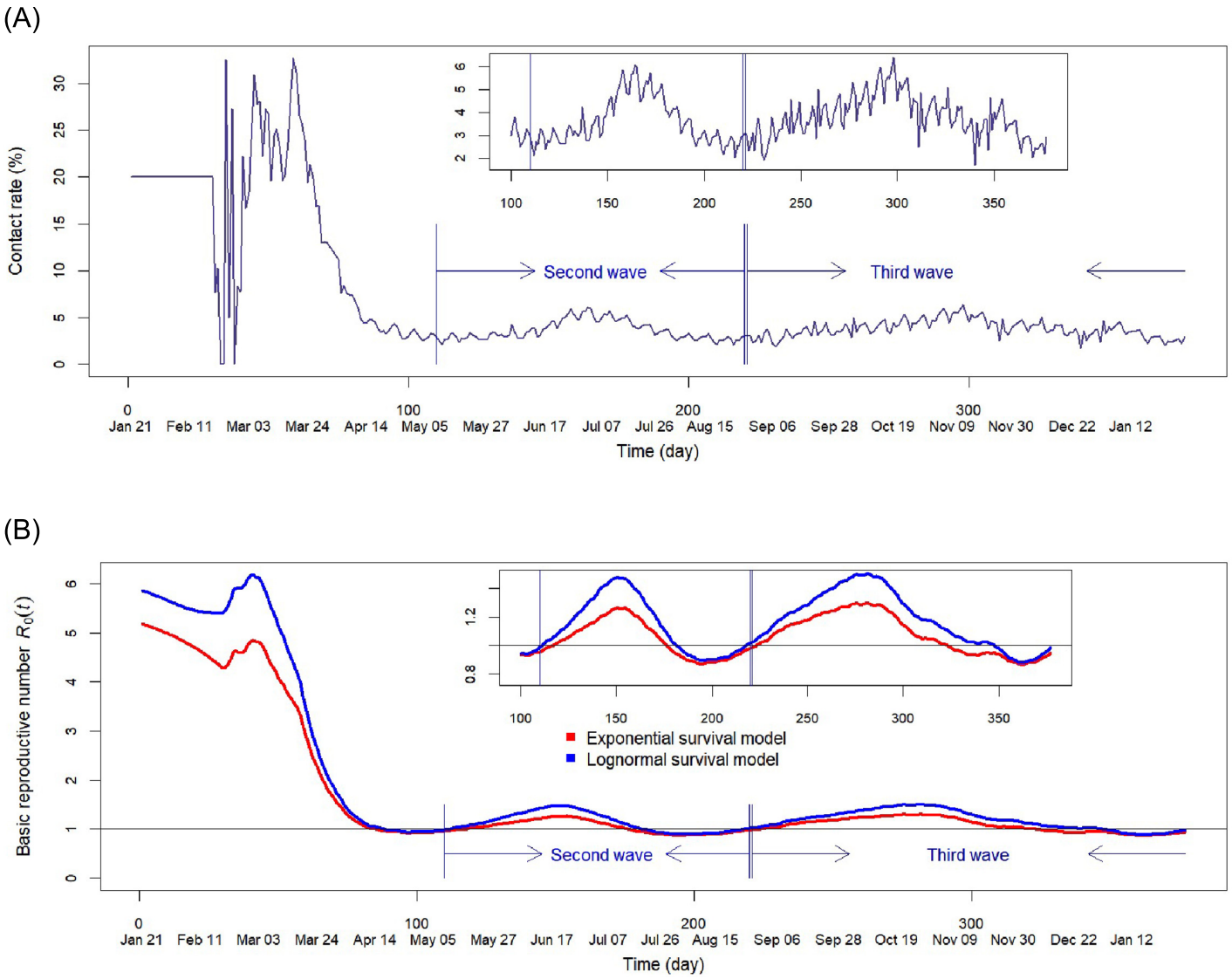

Fig 3. Pre-vaccination data. (A) Contact rate calculated as the fraction of new infected on day t in the total infected during day $t - 29$ through day t and set to 0.2 for the first 30 days. (B) Reproduction number $R_0(t)$ assuming data driven contact rate for exponential model (classical SIR model) and lognormal model where CFR is 8.5% for the first wave and 1.7% thereafter.

https://doi.org/10.1371/journal.pone.0254397.g003

The basic reproduction number, denoted as R_0, equals expected number of infections arising due to contact with a positive case in a population where all individuals are susceptible to infection [38–40]. With a time varying contact rate, the reproduction number generalizes to

$$R_0(t) = \int_0^\infty \left[\int_t^{t+x} \beta(s)\, ds \right] f_X(x)\, dx$$

which equals total (integrated) contact rate for an individual infected at t, averaged with respect to length of infectious duration, X. A discrete estimate for $R_0(t)$ was obtained by discrete integral calculus over β_t where the density f_X for X is obtained from the discrete survival model (Section S1.3 in S1 Appendix). Note if contact rate is constant $\beta_t = \beta$, then $R_0 = \beta\bar{X}$ which equals β/γ for the classical compartmental model [8, 10, 17, 41].

Results

Survival model specification using a simulation with fixed contact rate

We first examined the effect of survival model specification on disease spread using a simulation under a fixed contact rate. Using the discrete time algorithm, we compared performance of exponential and lognormal distributed models (specified in S1 and S2 Figs in S1 Appendix). The contact rate was set to a constant $\beta = 0.2$. Survival models were calibrated to have equal infectious duration and CFR. We observe significantly different behavior for the models. For the lognormal, peak of daily deaths occurs after peak of infectious cases, while for the exponential, peaks occur at the same time. This delay pattern for the lognormal is more realistic. Death and infectious peaks are highest for the lognormal (S3 and S4 Figs in S1 Appendix; S1 Video). Therefore, even with the same mean infectious duration and CFR, the type of survival model yields substantial difference in disease spread.

Analysis of pre-vaccination data

U.S. pre-vaccination data was then analyzed using the fully time varying SICD model, which included the time-varying data-driven discrete time contact rate β_t described earlier. The latter is shown in Fig 3A. Models were fit with CFR set to 8.5% so as to generate realistic proportion of daily deaths over daily infected. All models are able to reasonably approximate aCFR for the first wave defined as COVID-19 prior to May 9th (S5 Fig in S1 Appendix). However, all models overestimate aCDR after first wave.

Given this overestimation, we hypothesized that CFR must have decreased after easing of lockdown measures. To investigate this, models were re-estimated under previous parameter values but assuming a decreased CFR of 1.7% for the period following the first wave. To estimate values, the discrete time algorithm was applied to data in the period defined by the first wave using a CFR of 8.5% and then separately to post-first wave data using a CFR of 1.7%.

Fig 3B displays estimated $R_0(t)$ under the above settings. Both lognormal and exponential models have $R_0(t)$ that begin approximately at 1.0 at the start of second wave. Values increase and decrease completing a full period returning to the starting value of 1.0. This provides further confirmation of a second wave distinct from previous values (a similar pattern is observed for the third wave although it has a longer period). Under this adjusted decreased CFR, the lognormal model is now able to accurately approximate observed values of aCFR, daily infected and deaths, over all periods of the data (Fig 4F, S2 Video). The exponential model (Fig 4C) is however unreliable and underestimates both number of daily deaths and infectious cases. This is due to the faster recovery rate imposed by the exponential distribution assumption. A bimodal lognormal distribution included in our comparison also performs poorly (S6C Fig in S1 Appendix). From Fig 4 it can be concluded the lognormal model is the most accurate and realistic model. Only this model will be considered for the remaining analysis.

What-if analysis: Practical identifiability

Values for the bivariate process of daily new infected cases and daily deaths, denoted by $(\dot{I}(t), \dot{D}(t))$, were estimated under different parameters for the SICD lognormal model as a means to assess practical identifiability. The lognormal model is dependent on four parameters, and these are tuned according to desired values for $\bar{X}$ and M_{death}. For this reason, the structural parameter of interest θ for practical identifiability can be considered to be $(\bar{X}, M_{\text{death}})$. Thus practical identifiability for the SICD model is assessed by considering $g(t;\theta) = (\dot{I}(t), \dot{D}(t); \theta)$. Practical identifiability means that $g(t;\theta)$, number of daily new

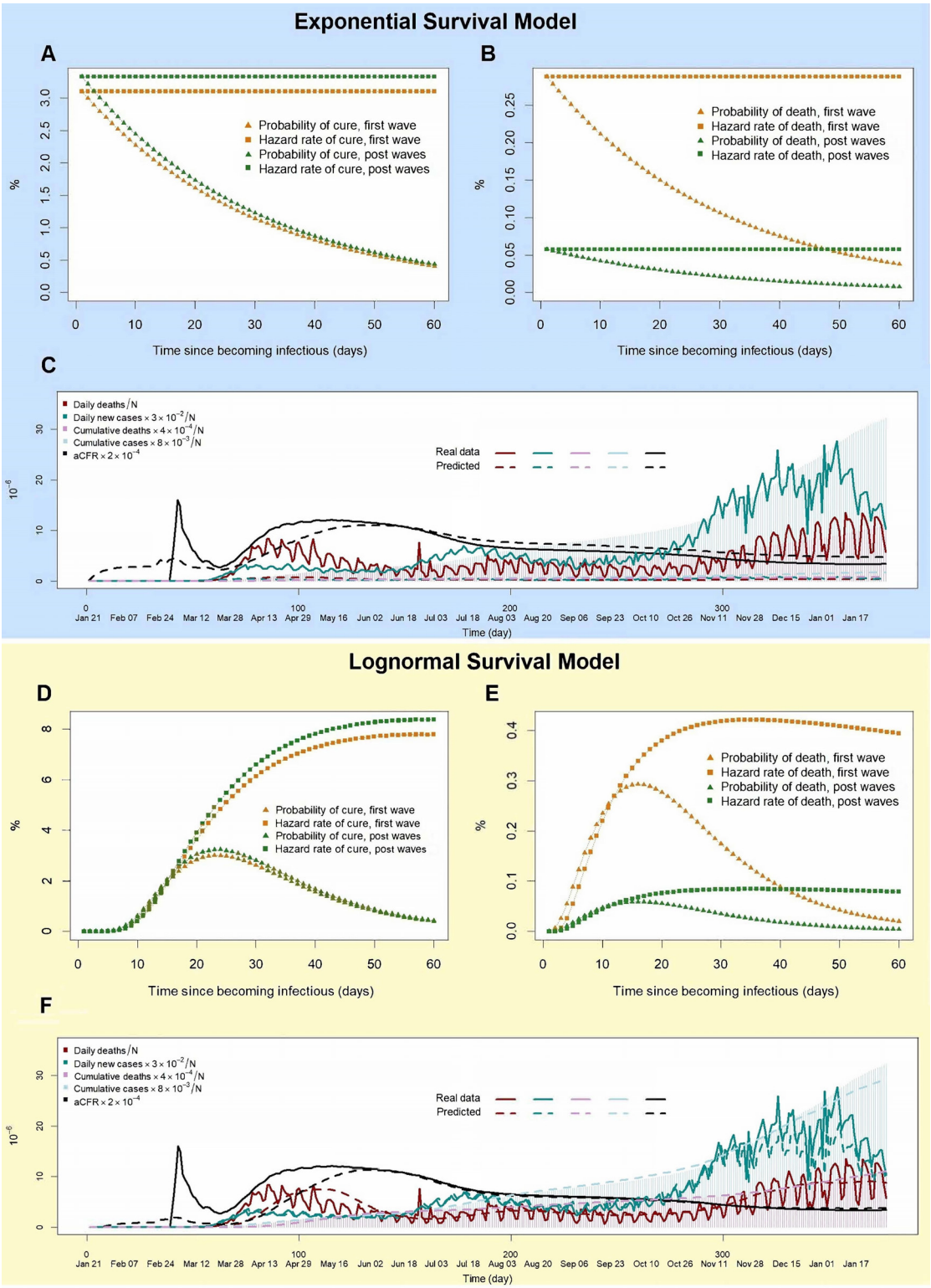

Fig 4. COVID-19 analysis of pre-vaccination data using exponential model (classical SIR model) and lognormal model assuming CFR is 8.5% for the first wave and 1.7% thereafter. (A,B) Probability and hazard values for cure and death for exponential model. (C) Estimated aCFR, cumulative and daily infected cases using exponential model compared to observed values. (D,E) Probability and hazard values for cure and death for lognormal model. (F) Estimated aCFR, cumulative and daily infected cases using lognormal model compared to observed values.

https://doi.org/10.1371/journal.pone.0254397.g004

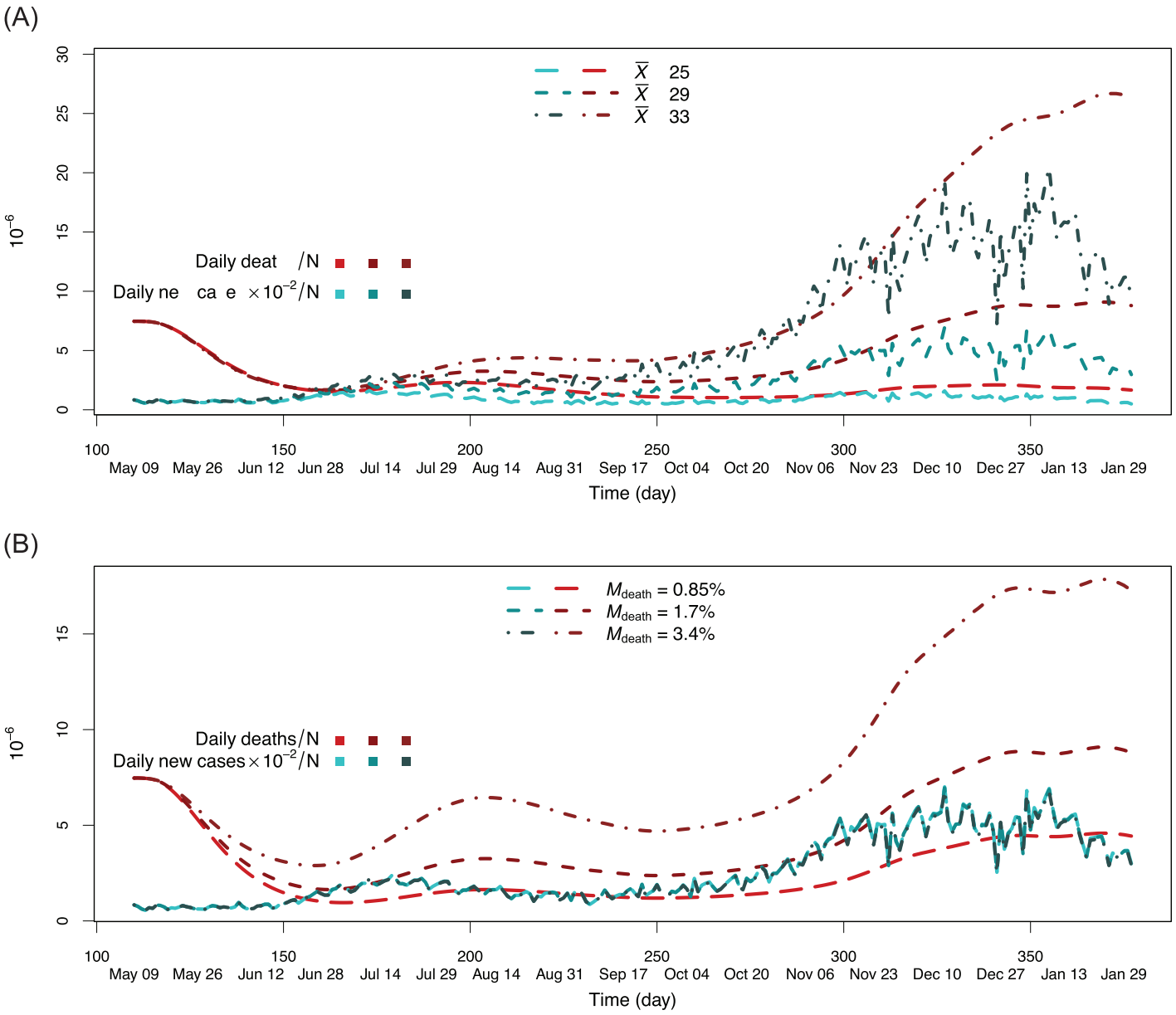

Fig 5. What-if practical identifiability analysis of SICD lognormal model. Estimated values for daily infected and deaths are given by for different $\bar{X}$ and M_{death} values (i.e. what if $\bar{X}$ equals this and what if M_{death} equals this): (A) $\bar{X} = 25, 29, 33$ (B) M_{death} = 0.85%, 1.7% and 3.4%.

https://doi.org/10.1371/journal.pone.0254397.g005

infected cases and daily deaths, is identified as a function of θ, i.e. mean infectious duration and CFR.

Fig 5(A) displays estimated daily new infected cases and daily deaths under different settings for $\bar{X}$. Three values were used, $\bar{X} = 25, 29, 33$, with all other parameters for the SICD lognormal model set to previous values. Fig 5(B) displays estimated values using M_{death} = 0.85%, 1.7% and 3.4%. All other parameters were set as before. In both figures at time zero, $g(t; \theta)$ is the same, however as t increases the bivariate process $g(t, \theta)$ is identified in θ. In particular, note that in Fig 5(B) although number of daily new cases is not affected by varying M_{death} (which is to be expected), the number of daily deaths changes quite dramatically. Thus when taken together as a bivariate proceess, this shows $g(t;\theta)$ is identified.

This what-if analysis also revealed the importance of the two parameters studied. To put some numbers to this, assuming $\bar{X} = 29$ days, we estimated 23,914,491 cumulative infected cases with 18,848,443 of these being cured after 377 days. If infectious time is shortened so that $\bar{X} = 25$, we estimate only 9,326,640 cumulative infected case of which 8,357,707 are cured. This also results in the number of cumulative deaths being reduced from 452,522 to 248,085. This demonstrates the importance cure time has on the disease.

2021 vaccination data and extension to a vaccine compartment

Although from Fig 4 the estimated deaths and aCFR are very close to true recorded numbers for 2020, estimated infected cases in early 2021 were found to be underestimated. There is a second peak around January 6th, 2021 with the largest number of daily new infected cases. This is likely due to the especially high contact rate during the holidays.

Therefore, between December 20th, 2020 and January 5th, 2021, the contact rate was increased to $\beta_t \leftarrow 1.35 \times \beta_t$ $(335 \leq t \leq 351)$ which allows the estimated infected cases to match the observed cases. However, deaths after 2021 are overestimated, which indicates that vaccines that became available in December 2020 must have improved survival. Therefore to address this, we extended the SICD model to include a vaccination compartment. This new model is referred to as the Susceptible Infectious Vaccinated Cure Death Immune (SIVCDI) model.

In this extension, the suspectible and infectious are separated into two groups. Unvaccinated susceptible cases are denoted by S^U and vaccinated susceptible cases are denoted by S^V, with $S = S^U + S^V$ as their sum. Likewise, unvaccinated infectious cases are denoted as I^U and vaccinated infectious are denoted as I^V, with $I = I^U + I^V$ as their sum. Associated parameters are also separated into two groups. The SIVCDI model is as follows (see Fig 6(A)):

$$\begin{aligned}
\frac{dS^U}{dt} &= -\alpha(t)S^U - \frac{\beta^U(t)IS^U}{N} \\
\frac{dS^V}{dt} &= \alpha(t)S^U - \frac{\beta^V(t)IS^V}{N} - \eta(t)S^V \\
\frac{dI^U}{dt} &= \frac{\beta^U(t)IS^U}{N} - \gamma^U(t)I^U \\
\frac{dI^V}{dt} &= \frac{\beta^V(t)IS^V}{N} - \gamma^V(t)I^V \\
\frac{dR}{dt} &= \gamma^U(t)I^U + \gamma^V(t)I^V + \eta(t)S^V.
\end{aligned}$$

Here $\alpha(t)$ is the vaccination rate at time t, β^U is the effective contact rate for unvaccinated cases at time t, equal to average number of contacts per person per time multiplied by probability of disease transmission between a unvaccinated susceptible and infectious case at time t, β^V is the contact rate between a vaccinated susceptible case and infectious case at time t, $\gamma^U(t)$ denotes the removal rate for unvaccinated infectious cases at time t, $\gamma^V(t)$ is the removal rate for the vaccinated infectious cases and $\eta(t)$ is the immune rate equal to percentage of vaccinated individuals who become immune to the disease. As before, the sample size N is fixed and the above equations can be reduced by one using $N = S^U + S^V + I + R$.

The analysis used 2020 and 2021 data ($t \geq 335$) from the New York Times COVID-19 repository [32], combined with publicly available vaccination data [42] recording number of vaccines administered per day. Calculations were based on a discrete time algorithm (Section S4.1 in S1 Appendix). Previously for the SICD model, X equals the continuous event time of an infected individual who either recovers from infection or dies due to infection. With the SIVCDI model, X becomes X_U and we add a new continuous variable X_V, defined as the event

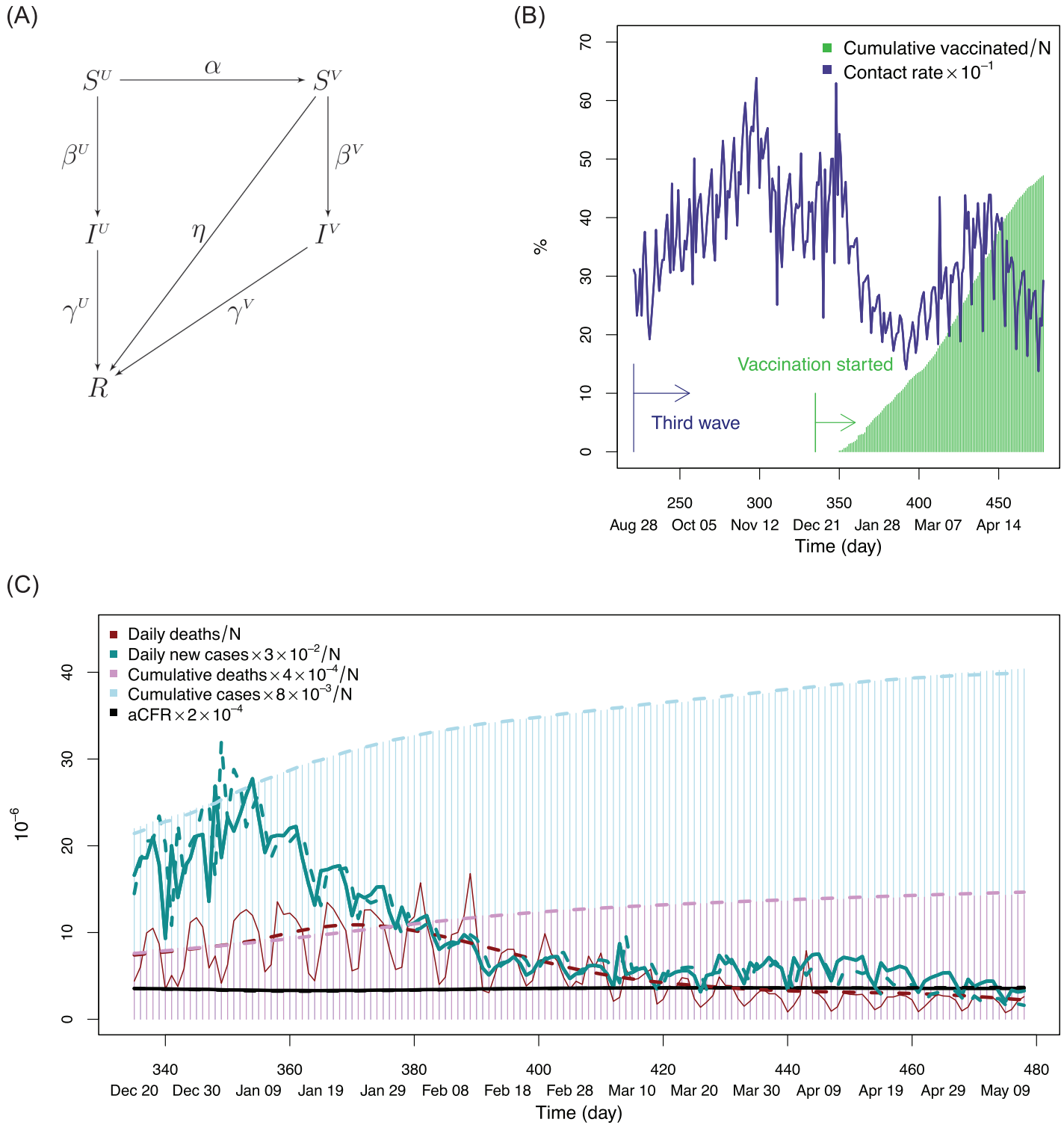

Fig 6. Vaccination model results. (A) Graphical representation of SIVCDI model. Removed status includes cured, dead, and immune from vaccination. (B) Contact rate and percentage of vaccinated population. Contact rate shows presence of a small wave after February 2021, which could be a potential fourth wave, but because of the success of vaccination, the fourth wave contact rate is much lower than the third wave contact rate. (C) Results of the data driven SIVCDI model from December 20th, 2020 to May 11th, 2021, where dashed lines are predicted values and solid lines are observed values.

https://doi.org/10.1371/journal.pone.0254397.g006

time that an infected vaccinated individual either recovers or dies. Related to X_V is is another new variable E_V, defined as the event time for a vaccinated individual who becomes immune. We define the mean immunity period as $\bar{E}_V$, equal to average number of days vaccinated individuals who become immune require to develop immunity. This was set to $\bar{E}_V = 30$ days but results were fairly robust to its choice. Based on the previous analysis showing benefit of improved infectious time, the average cure time for infected vaccinated individuals was set to 25 days. Mortality rates for vaccinated individuals has been observed to be extremely low, therefore the death rate for this group was set to $M_{\text{death}} = .001\%$. Contact rates β^U and β^V were estimated using data from published studies of vaccine efficacy (Section S4.2 in S1 Appendix). All other parameters including parameters for the unvaccinated infecious cases were set as before for the pre-vaccination data analysis. The results for the data driven SIVCDI model are given in Fig 6(C). As can be seen, predicted values are near identical to observed values. The contact rate is displayed in Fig 6(B) and shows a reduced wave after February 2021, which could be a potential fourth wave. Its value is reduced relative to the third contact rate wave due to success of the vaccination.

Discussion

Time to event subject to competing risks is a branch of survival analysis called competing risk analysis. While clinical data is often used in medicine for such analyses, this type of setting is very different than the setting where a general population is exposed to a rapidly manifesting infectious disease. For clinical data, the initial time is often recorded as time of hospitalization, or diagnosis for a specific stage of disease development [43, 44], rather than onset of disease, which is what is needed to model infectious duration of a disease. Also, clinical settings often target specific populations and therefore their results may not translate to general populations. With this in mind, we developed our compartmental model using a competing risk framework (SICD model) where survival model parameters can be determined from aggregated population level epidemiological data. This is the type of data one typically has to work with when an infectious disease strikes.

The advantages of a flexible competing risk framework are clear when compared with the classical SIR model which assumes a constant hazard. In infectious diseases, the hazard rate at the beginning of disease development is often very low with increasing values later in time. The lognormal distribution [45, 46] is well suited for this type of modeling as it accomodates hazards that can increase and decrease [47]. The distribution is conveniently specified by two parameters, and these can easily be tuned so that peak of death occurs after peak of infectious cases, which is suggested by COVID-19 data. Separate lognormal parameter values are used for cure and death as the magnitude of risk for cure is much larger than death.

This work has shown using just four parameters, that the lognormal distribution can accurately model COVID-19 pre-vaccination data when used in combination with a data-driven dynamic contact rate. The model was able to accurately estimate dynamic values like the aCFR and daily number of infected and deaths, but also at the same time provide estimated values for key survival parameters such as the CFR. Although the lognormal was used exclusively here, other distributions could also be used; for example, a promising choice might be the Erlang distribution [30]. However the lognormal proved robust in our experimentation. Also because of certain numerical simplifications that occur for this distribution (Section S2.1 in S1 Appendix), we found it very convenient for numerical calculations.

The SICD competing risk model was extended by adding a vaccination compartment and applied to 2021 vaccination data. Parameters for the extended SIVCDI model were separated into the two groups of vaccinated and unvaccinated and were relatively easy to specify using

published studies and publicly available data. As was shown, the SIVCDI model could accurately fit the 2021 observed data. As has been noted, parameters that played a crucial role in this accurate modeling were a lower cure time and a lower mortality rate for those vaccinated. Both adjustments are quite realistic given the importance of a reduced cure time suggested by a what-if analysis and the wide spread consensus that vaccination signficantly reduces mortality.

Supporting information

S1 Appendix.
(ZIP)

S1 Video. Comparison of three survival models with 29-day mean infectious duration, 8.5% morality and 0.2 fixed contact rate.
(MP4)

S2 Video. Analysis of COVID-19 U.S. data with with 29-day mean infectious duration and data-driven contact rate.
(MP4)

S1 Fig. Comparison of three different survival models where all models have identical mean infectious period $\bar{X} = 29$ and mortality rate M_{death} = 8.5%. Shown are CCDF $\bar{F}(t)$ (black), CIF for cure $F_1(t)$ (orange) and CIF for death $F_2(t)$ (green). (A) Scenario I uses an exponential distribution, which is equivalent to the classical SIR model by Corollary 1 of Appendix. (B) Scenario II uses a lognormal distribution. (C) Scenario III uses a bimodal lognormal distribution.
(PDF)

S2 Fig. Discrete time survival values for scenario I (red), II (blue) and III (purple). (A) Discrete time pseudo-densities for cure. Most infections recover at the beginning in scenario I; around 15-40 days in scenario II; and either within 17 days, or around 25 to 55 days, in scenario III. (B) Discrete time pseudo-densities for death. Most deaths occur at the beginning in scenario I and around 15-40 days in scenarios II and III. (C) Discrete time hazard rates for cure. Scenario I has constant hazard whereas scenarios II and III assume hazards that initially increase and then decrease. Scenario III assumes a bimodal shape for the cure hazard. (D) Discrete time hazard rates for death.
(PDF)

S3 Fig. Discrete time SIR models assuming a constant contact rate $\beta(t)$ = 0.2. Infectious cases $I(t)$ (black), daily cured cases $\dot{C}(t)$ (orange) and daily deaths $\dot{D}(t)$ (green) are displayed as percentage of total population. Daily cured and deaths being much smaller than $I(t)$ are multiplied by 30 and 200. (A) In scenario I, all values have the same trend and peak at the same time. (B) In scenario II, daily deaths peak after infectious cases, which is more realistic. (C) In scenario III, deaths also peak after infectious cases, but daily cured has two waves due to the bimodal distribution assumption.
(PDF)

S4 Fig. Comparison of discrete time SIR models assuming a constant contact rate $\beta(t)$ = 0.2. (A) Daily cured. (B) Daily deaths. (C) aCDR. (D) Infectious cases as percentage of population, $I(t)/N$. Values of aCDR should be very low at onset of disease due to few cures and death occuring immediately after infection. Therefore, scenario I is unrealistic.
(PDF)

S5 Fig. Analysis of COVID-19 pre-vaccination data assuming a constant mortality rate for first and subsequent waves. Scenario II is best at estimating daily new cases. However, aCDR and daily deaths are overestimated after first wave, thus suggesting a lower mortality for post-first wave data.
(PDf)

S6 Fig. Analysis of COVID-19 pre-vaccination data assuming a lower mortality rate for second and subsequent waves. (A) Basic reproduction number $R_0(t)$; note its values are much smaller for Scenarios I and III than II. (B) Even though I and III have similar $R_0(t)$ profiles, estimated values for daily new infections and deaths are different. (C) Bimodal lognormal distribution continues to perform poorly even under assumption of lower mortality for post-first wave data.
(ZIP)

Author Contributions

Conceptualization: Min Lu, Hemant Ishwaran.

Data curation: Min Lu, Hemant Ishwaran.

Formal analysis: Min Lu, Hemant Ishwaran.

Methodology: Min Lu, Hemant Ishwaran.

Writing – original draft: Min Lu, Hemant Ishwaran.

Writing – review & editing: Min Lu, Hemant Ishwaran.

References

1. Wu Z, McGoogan J. Outbreak in China: Summary of a Report of 72314 Cases from the Chinese Center for Disease Control and Prevention. JAMA. 2020; 10.
2. Ngonghala CN, Iboi E, Gumel AB. Could masks curtail the post-lockdown resurgence of COVID-19 in the US? Mathematical Biosciences. 2020; p. 108452.
3. Li HL, Jecker NS, Chung RYN. Reopening economies during the COIVD-19 pandemic: reasoning about value tradeoffs. The American Journal of Bioethics. 2020; 20(7):136–138. https://doi.org/10.1080/15265161.2020.1779406
4. Shear MD, Against MSTEP. Governors Who Have Imposed Virus Restrictions. New York Times April. 2020; 17.
5. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. The Lancet infectious diseases. 2020;. https://doi.org/10.1016/S1473-3099(20)30195-X PMID: 32171390
6. Spychalski P, Blażyńska-Spychalska A, Kobiela J. Estimating case fatality rates of COVID-19. The Lancet Infectious Diseases. 2020;. https://doi.org/10.1016/S1473-3099(20)30246-2 PMID: 32243815
7. Mantha S. Ratio, rate, or risk? The Lancet Infectious Diseases. 2020;. https://doi.org/10.1016/S1473-3099(20)30439-4 PMID: 32473091
8. Anderson RM, Anderson B, May RM. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press; 1992.
9. Diekmann O, Heesterbeek H, Britton T. Mathematical Tools for Understanding Infectious Disease Dynamics. vol. 7. Princeton University Press; 2012.
10. Hethcote HW. The mathematics of infectious diseases. SIAM review. 2000; 42(4):599–653. https://doi.org/10.1137/S0036144500371907
11. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Princeton University Press; 2011.
12. Brauer F, Castillo-Chavez C, Castillo-Chavez C. Mathematical Models in Population Biology and Epidemiology. vol. 2. Springer; 2012.

13. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health. 2020;. https://doi.org/10.1016/S2214-109X(20)30074-7 PMID: 32119825

14. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases. 2020;. https://doi.org/10.1016/S1473-3099(20)30144-4 PMID: 32171059

15. Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of The Total Environment. 2020; p. 138817.

16. Kufel T. ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries. Equilibrium Quarterly Journal of Economics and Economic Policy. 2020; 15(2):181–204.

17. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1927; 115(772):700–721.

18. Qianying L, et al. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan. China with Individual Reaction and Governmental Action Int J Infect Dis. 2020; 93:211–216.

19. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLOS One. 2020; 15(3):e0230405. https://doi.org/10.1371/journal.pone.0230405

20. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health. 2020;. https://doi.org/10.1016/S2468-2667(20)30073-6 PMID: 32220655

21. Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nature Medicine. 2020; p. 1–6. https://doi.org/10.1038/s41591-020-0883-7 PMID: 32322102

22. Lofgren ET, Moehring RW, Anderson DJ, Weber DJ, Fefferman NH. A mathematical model to evaluate the routine use of fecal microbiota transplantation to prevent incident and recurrent Clostridium difficile infection. Infection control and hospital epidemiology: the official journal of the Society of Hospital Epidemiologists of America. 2014; 35(1):18. https://doi.org/10.1086/674394

23. Hewitt J, Carter B, Vilches-Moraga A, Quinn TJ, Braude P, Verduri A, et al. The effect of frailty on survival in patients with COVID-19 (COPE): a multicentre, European, observational cohort study. The Lancet Public Health. 2020;. https://doi.org/10.1016/S2468-2667(20)30146-8 PMID: 32619408

24. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. JAMA. 2020;. https://doi.org/10.1001/jama.2020.6775 PMID: 32320003

25. Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. PNAS. 2020; 117(29):16732–16738. https://doi.org/10.1073/pnas.2006520117

26. Lloyd AL. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. Theoretical population biology. 2001; 60(1):59–71. https://doi.org/10.1006/tpbi.2001.1525

27. Hethcote HW, Tudor DW. Integral equation models for endemic infectious diseases. Journal of Mathematical Biology. 1980; 9(1):37–47. https://doi.org/10.1007/BF00276034

28. Keeling MJ, Grenfell BT. Disease extinction and community size: modeling the persistence of measles. Science. 1997; 275(5296):65–67. https://doi.org/10.1126/science.275.5296.65

29. Lloyd AL. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. Proceedings of the Royal Society of London Series B: Biological Sciences. 2001; 268(1470):985–993. https://doi.org/10.1098/rspb.2001.1599

30. Krylova O, Earn DJ. Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. Journal of The Royal Society Interface. 2013; 10(84):20130098. https://doi.org/10.1098/rsif.2013.0098

31. Champredon D, Dushoff J, Earn DJ. Equivalence of the Erlang-distributed SEIR epidemic model and the renewal equation. SIAM Journal on Applied Mathematics. 2018; 78(6):3258–3278. https://doi.org/10.1137/18M1186411

32. Times NY. New York Times COVID-19 Data for the US, 2020; accessed December 27, 2020. Available from: https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv.

33. Bellomo N, Bingham R, Chaplain MA, Dosi G, Forni G, Knopoff DA, et al. A multi-scale model of virus pandemic: Heterogeneous interactive entities in a globally connected world. arXiv preprint arXiv:200603915. 2020;.

34. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application.

Annals of Internal Medicine. 2020; 172(9):577–582. https://doi.org/10.7326/M20-0504 PMID: 32150748

35. Ng Y, Li Z, Chua YX, Chaw WL, Zhao Z, Er B, et al. Evaluation of the effectiveness of surveillance and containment measures for the first 100 patients with COVID-19 in Singapore–January 2–February 29, 2020. Morbidity and Mortality Weekly Report. 2020; 69:307–311. https://doi.org/10.15585/mmwr.mm6911e1 PMID: 32191691

36. Tuncer N, Le TT. Structural and practical identifiability analysis of outbreak models. Mathematical biosciences. 2018; 299:1–18. https://doi.org/10.1016/j.mbs.2018.02.004

37. Massonis G, Banga JR, Villaverde AF. Structural identifiability and observability of compartmental models of the COVID-19 pandemic. Annual reviews in control. 2020;. https://doi.org/10.1016/j.arcontrol.2020.12.001 PMID: 33362427

38. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. Science. 2009; 324(5934):1557–1561. https://doi.org/10.1126/science.1176062 PMID: 19433588

39. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the basic reproduction number (R0). Emerging Infectious Diseases. 2019; 25(1):1. https://doi.org/10.3201/eid2501.171901

40. Guerra FM, Bolotin S, Lim G, Heffernan J, Deeks SL, Li Y, et al. The basic reproduction number (R0) of measles: a systematic review. The Lancet Infectious Diseases. 2017; 17(12):e420–e428. https://doi.org/10.1016/S1473-3099(17)30307-9 PMID: 28757186

41. Diekmann O, Heesterbeek J. Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis, and Interpretation. John Wiley and Sons, Inc., New York; 2000.

42. Hannah R, Esteban OO, Diana B, Edouard M, Joe H, Bobbie M, et al. Coronavirus Pandemic (COVID-19). Our World in Data. 2020;.

43. Crowley J, Hu M. Covariance analysis of heart transplant survival data. Journal of the American Statistical Association. 1977; 72(357):27–36. https://doi.org/10.1080/01621459.1977.10479903

44. Collett D. Modelling Survival Data in Medical Research. CRC press; 2015.

45. Chapman J, O'callaghan C, Hu N, Ding K, Yothers G, Catalano P, et al. Innovative estimation of survival using log-normal survival modelling on ACCENT database. British Journal of Cancer. 2013; 108(4):784–790. https://doi.org/10.1038/bjc.2013.34 PMID: 23385733

46. Royston P. The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. Statistica Neerlandica. 2001; 55(1):89–104. https://doi.org/10.1111/1467-9574.00158

47. Aragao GM, Corradini MG, Normand MD, Peleg M. Evaluation of the Weibull and log normal distribution functions as survival models of Escherichia coli under isothermal and non isothermal conditions. International Journal of Food Microbiology. 2007; 119(3):243–257. https://doi.org/10.1016/j.ijfoodmicro.2007.08.004