The 1978 English boarding school influenza outbreak: where the classic SEIR model fails

Konstantin K. AVILOV[1*], Qiong LI[2], Lixin LIN[3], Haydar DEMIRHAN[3], Lewi STONE[3,4], Daihai HE[1*]

1 Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong SAR, China

2 Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, Zhuhai, China

3 Mathematical Sciences, School of Science, Royal Melbourne Institute of Technology (RMIT) University, Melbourne, Victoria, Australia

4 Biomathematics Unit, Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel

Abstract

Previous work has failed to fit classic SEIR epidemic models satisfactorily to the prevalence data of the famous English boarding school 1978 influenza A/H1N1 outbreak during the children's pandemic. It is still an open question whether a biologically plausible model can fit the prevalence time series and the attack rate correctly. To construct the final model, we first used an intentionally very flexible and overfitted discrete-time epidemiologic model to learn the epidemiological features from the data. The final model was a susceptible (*S*) - exposed (*E*) - infectious (*I*) - confined to bed (*B*) - convalescent (*C*) - recovered (*R*) model with time delay (constant residence time) in *E* and *I* compartments and multistage (Erlang-distributed residence time) in *B* and *C* compartments. We simultaneously fitted the reported *B* and *C* prevalence curves as well as the attack rate (proportion of children infected during the outbreak). The non-exponential residence times were crucial for good fits. The estimates of the generation time and the basic reproductive number ($\mathcal{R}_0$) were biologically reasonable. A simplified discrete-time model was built and fitted using the Bayesian procedure. Our work not only provided an answer to the open question, but also demonstrated an approach to constructive model-generation.

Keywords modelling, children's pandemic, delay differential equations, residence time, influenza progression model, Bayesian epidemic model.

[*] Corresponding authors: konstantin.avilov@polyu.edu.hk, kkavilov@gmail.com, daihai.he@polyu.edu.hk

Introduction

Mathematical models have long been used as tools for providing plausible explanations of the epidemiologic mechanisms underlying respiratory virus outbreaks, both city-level and community-level. For example:

- Simple differential equation models have been successfully used to explain the multiple death waves in 1918 influenza A/H1N1 pandemic at the city level [1, 2].
- Such models were also used in isolated small size populations (community level) for two waves of cases due to reinfection for A/H3N2 [3].
- The COVID-19 outbreak onboard the Diamond Princess cruise ship, 2020 [4] is a widely used and important modelling example of an outbreak in a closed population.

Our work deals with the unexpected and unusual return of influenza A/H1N1 in 1977 and, more specifically, with a dataset of the 1978 England boarding school influenza outbreak [5] that occurred during the 1977-1978 influenza A/H1N1 pandemic, or the so-called *children's pandemic* [5-8]. This dataset is a classical example used in numerous mathematical biology textbooks and lectures [8-12]. The data from a boarding school contains: 1) the daily populations of confined to bed schoolboys and convalescent schoolboys; 2) the total attack rate (the number of persons eventually infected): 512 out of 763 students were impacted (67%) [5].

The return of A/H1N1 was brought up in the discussion of the 2019 COVID-19 pandemic [13], although it had always been a topic of gain-of-function research before the pandemic [14]. Historically, influenza A/H1N1 first appeared in 1918 and was referred to as the Spanish flu, a pandemic that led to the deaths of >60 million people across the globe. A/H1N1 continued to circulate for 40 years, and was then replaced by H2N2 in 1957. It then reemerged in 1977 as H1N1/77 (also known as the Russian Flu), and co-circulated with H3N2/68, until 2009. Given that the H1N1/77 virus was almost identical to the main strain from that in the 1950's, there have been claims that the reappearance was due to a laboratory leak of a stored sample [14].

The boarding school outbreak of A/H1N1 in 1978 was explosive (67% were infected during ≈14 days) and there are records of many similar events. For example, "[t]he outbreak at the U.S. Air Force Academy (USAFA) was so severe — over the course of 9 days, 76%, or 3,280 cadets, became ill — that all academic and military training was suspended" [14]. More than 45 years have passed since the boarding school outbreak and, despite many attempts to fit the data with mathematical models, all attempts to date were either wrong or provided unsatisfactory fits [8-11, 15, 16]. For example, Prof. M.Y. Li [15] called this dataset "an epidemic enigma" and noted that previous studies had drastically overestimated the attack rate of the outbreak (and this problem is common in mathematical epidemiology – for example, in Ebola studies [17]). Kalachev et al. [16] pointed out that previous attempts [8] misinterpreted the data and forced them into the pure and overly simplistic SIR-type modelling paradigm. However, even Kalachev's improved and more realistic models were still problematical. Importantly, their estimated attack rate was close to 100%, and failed to match the reported 67%.

In this paper, we build a mathematical model that achieves all three fitting targets, which, to the best of our knowledge, has not been successfully achieved by other authors: (i) the time series of the daily number of children confined to bed, (ii) the time series of the daily number of convalescents, and (iii) the total attack rate. Our objective is to find a biologically plausible model which can fit the observed data satisfactorily – in particular, the attack rate, which was omitted so far in previous works. We hope that the revisit of the dated 1978 example will also help other researchers with the very general problem of overestimating the attack rate.

The boarding school dataset indicates how pandemic influenza is transmitted in a boarding school setting, and reminds us of similar occurrences in the 2019 COVID-19 pandemic, e.g. a cruise ship (the Diamond Princess cruise ship), nursing homes, or prisons [18]. Thus, it is still relevant and of significance to reveal the dynamics of the transmission of pandemic viruses in such a setting with biologically justifiable mathematical models.

The basic reproduction number and generation time

One of most important characteristics of the epidemic process is the basic reproduction number, $\mathcal{R}_0$, defined as "the expected number of secondary cases produced, in a completely susceptible population, by a typical infected individual" [19]. Another important quantity is the generation time (GT), defined as the time delay between the infection time of an infector and that of their infectee. Estimations of the basic reproduction number $\mathcal{R}_0$ and the distribution of GT are crucial both for understanding the general dynamics of the spread of the infection and for planning vaccination or quarantine campaigns.

Data

The source of the data on the outbreak in a boarding school for boys in England is an anonymous publication in the British Medical Journal [5]. It gives the daily prevalences only as points on the graph and therefore has to be digitised. Thus, there are slight variations in different datasets of the event. We used the data as presented in the R package "outbreaks" [20] – see Fig.1. and Appendix Section 1. The source publication [5] reports the total number of boys infected (512 persons) and the total number of schoolboys in the school (763 persons).

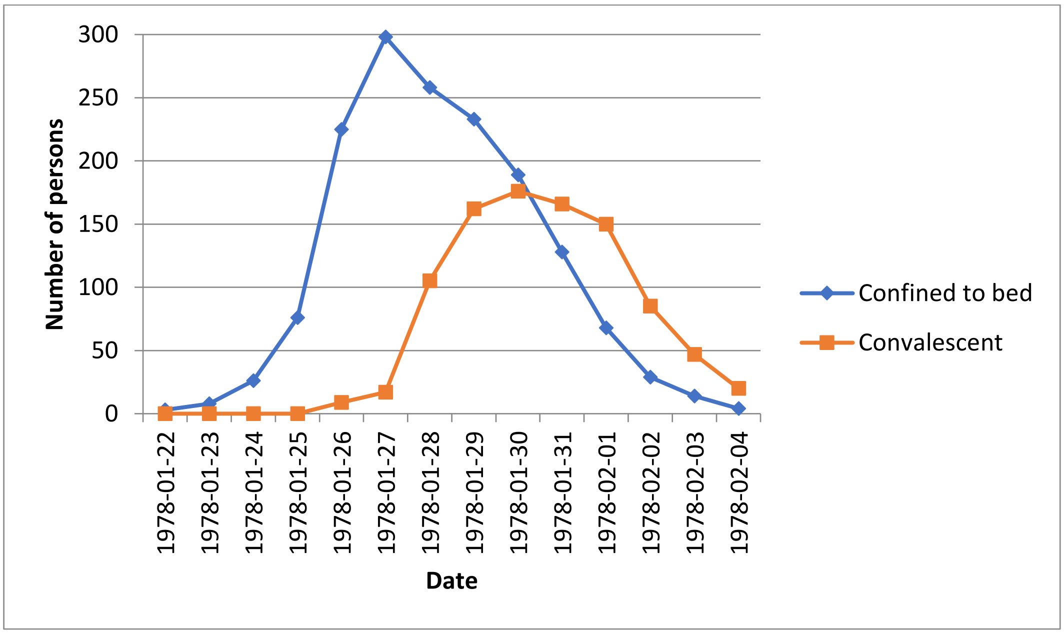

Figure 1. The daily numbers of confined to bed and convalescent schoolboys during the 1978 English boarding school influenza A/H1N1 outbreak; as presented in [20] (reproduced ultimately from [5]).

Previous SIR-like models

The boarding school dataset has been approximated with standard SIR or SEIR epidemiologic models numerous times (e.g., [8-12]). The data on attack rate (AR) and the number of convalescents were usually ignored, the number of confined to bed students was interpreted as the size on the *I*-group (infectious), and the goodness-of-fit was often mediocre.

To illustrate this point, we fitted a number of SIR/SEIR model variants (see the Appendix, Section 2 for full details and graphs):

- SIR and SEIR models with a completely naïve starting population ($R_{t=0} = 0$, the typical approach) fail to approximate the attack rate: they converge to a nearly-100% attack rate;
- SIR and SEIR models with an estimated, non-zero initial number of immune individuals ($R_{t=0}$) can approximate the correct attack rate, but the goodness-of-fit even in the *I*-curve alone is not satisfactory ($RMSE =$ 16–16.5 pers.; moreover, the time-span of the outbreak is significantly overestimated).

Prof. M.Y. Li [15] described it as “there is no known model of any kind in the literature that has correctly described both the time course of the epidemic ($I(t)$) and the final size”.
Kalachev et al. [16] believed it to be erroneous to interpret the “confined to bed” persons (*B*) as the “infected and infectious” (*I* in the classical *SEIR* model).

Methods

General modelling idea

In contrast to other studies, Kalachev et al. [16] argued that the *B*-state is actually a later stage of the disease preceded by an unobserved highly infectious state. This is also supported by the information from the initial publication [5] which mentioned that “symptoms subsided quickly once the boys were confined to bed”. Kalachev et al. called their approach “a very natural interpretation” and “the most realistic for the available data”.

Following their work, we initially assumed that the individuals proceeded through the following consecutive stages:

- *S* – susceptible (not infected, not infectious),
- *E* – exposed (infected, not infectious yet),
- *I* – infectious (active disease),
- *B* – confined to bed (still ill, potentially infectious (to be estimated)),
- *C* – convalescent (possibly ill, not infectious),
- *R* – recovered and/or immune (not ill, not infectious).

It is natural to assume that, on a two-week time interval, people’s behaviour and contact patterns changed little, so that all effects of quarantining can be captured by the difference between the *I*- and *B*-states. On longer time intervals (months or years), the infectivity rate would likely change. We therefore incorporated time-independent effective infectiousness (infectivity rate) into our models and attempted to explain all features of the data with an appropriate model of disease progression.

An important feature of our models was non-exponential distributions of residence times in some compartments. Classic linear ODE-based models like $\mathrm{d}X/\mathrm{d}t = -\sigma X$ result in an exponential residence time distribution in X, and this may be a limiting factor in reproducing real or realistic residence times in certain biological states.

Building the model

We built our final model in two stages: first, we created a very flexible discrete-time model capable of reproducing every possible discrete distribution of residence time by having it defined as a tabulated function, and then fitted the model to the data; second, we constructed a simpler final model (continuous time, delay differential equations) that used residence time distributions and other model features close to those observed in the fitted flexible model.

The flexible discrete-time model included groups *S*, *E*, *I*, *B*, *C*, *R*, with tabulated distributions of residence time for groups *E*, *I*, *B*, *C*. Each table prescribed the probability distribution of the number of time-steps (days) that an individual spends in the given group. The model allowed for the possibility that both *I* and *B* became infectious, and other flexible features (a full definition of the model is in the Appendix, Section 3). This model was fitted to the data (hence, the tabulated distributions of residence time were estimated), and the results so obtained (see the Appendix) drove our choice of the features of the much simpler final continuous-time model:

- the survival function of *E* in the flexible model which abruptly dropped at two points in time (Fig. AF5) was interpreted in the continuous model as some *E*-individuals having zero residence time (“E-bypass”) and some having a Dirac-δ-distributed residence time in *E* (i.e., spending exactly the same time in the group); (survival function is the probability for an individual to remain in the given model group for at least the given amount of days, it is equal to $1 - CDF_\tau$, where CDF_τ is the cumulative distribution function of residence time τ)
- the abrupt drop in the survival function of *I* (Fig. AF3) called for a Dirac-δ-distributed residence time in *I* in the continuous model;
- the survival functions of *B* and *C* with little to no drop in the first days (Fig. AF3) called for non-exponential residence time distributions separated from zero (i.e., with a low density near zero);
- *B*-individuals’ infectivity was estimated to be zero in the flexible model (Table AT3), and this was postulated to be so in the final model;
- non-zero initial number of immune individuals $R_{t=0} > 0$ (Table AT3).

Final DDE model

The final model used the same groups (*S*, *E*, *I*, *B*, *C*, and *R*), but was defined with delay differential equations (DDEs). The features of the final SEIBCR model, based on the results of the flexible model, are:

- the residence time in *E*- and *I*-stages is Dirac-δ-distributed, and it is naturally modelled with DDEs;
- a fraction of newly infected individuals (p_{noE}, which is a parameter) “bypasses” the *E*-stage and goes directly to the *I*-stage, thus approximating the overall residence time distribution in *E* observed in the flexible model (Fig. AF5);

- the residence time in *B* and *C* follows the Erlang distribution, and is modelled by creating 10 consecutive dummy groups B_j and C_j, *j*=1,…,10, with equal-rate linear transfer terms between them (known as the "linear chain trick");
- the only source of infection is *I*-individuals; stages *B* and *C* are non-infectious; only *S*-individuals can be infected; there is no immunity loss;
- the population is fixed (no deaths, births, or migration).

The model consists of equations (1), and the variables and parameters are defined in Table 1.

$$
\begin{aligned}
&\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta_I \frac{SI}{N} \\
&\frac{\mathrm{d}E}{\mathrm{d}t} = (1-p_{noE})\beta_I \left(\frac{SI}{N} - \frac{S(t-T_E)I(t-T_E)}{N}\right) \\
&\frac{\mathrm{d}I}{\mathrm{d}t} = (1-p_{noE})\beta_I \left(\frac{S(t-T_E)I(t-T_E)}{N} - \frac{S(t-T_E-T_I)I(t-T_E-T_I)}{N}\right) \\
&+p_{noE}\beta_I \left(\frac{SI}{N} - \frac{S(t-T_I)I(t-T_I)}{N}\right) \\
&\frac{\mathrm{d}B_1}{\mathrm{d}t} = (1-p_{noE})\beta_I \left(\frac{S(t-T_E-T_I)I(t-T_E-T_I)}{N}\right) \\
&+p_{noE}\beta_I \left(\frac{S(t-T_I)I(t-T_I)}{N}\right) - N_{gB}\delta B_1 \\
&\frac{\mathrm{d}B_j}{\mathrm{d}t} = N_{gB}\delta\left(B_{j-1} - B_j\right), \qquad j = 2, \dots, N_{gB} \\
&\frac{\mathrm{d}C_1}{\mathrm{d}t} = N_{gB}\delta B_{N_{gB}} - N_{gC}\varepsilon C_1 \\
&\frac{\mathrm{d}C_j}{\mathrm{d}t} = N_{gC}\varepsilon\left(C_{j-1} - C_j\right), \qquad j = 2, \dots, N_{gC} \\
&\frac{\mathrm{d}R}{\mathrm{d}t} = N_{gC}\varepsilon C_{N_{gC}} \\
&[S, E, I, B_1, \dots, C_1, \dots, R](t=0) \\
&\qquad\qquad = [N - E_{t=0} - R_{t=0}, (1-p_{noE})E_{t=0}, p_{noE}E_{t=0}, 0, \dots, 0, \dots, R_{t=0}]
\end{aligned}
\tag{1}
$$

The lagged terms $\beta_I \frac{S(t-T)I(t-T)}{N}$ (where T can be T_E, T_I, or $T_E + T_I$) were assumed to be equal to 0 when $(t-T) < -0.25$ days, and equal to $E_0/0.25$ when $-0.25 \leq (t-T) < 0$ in order to model the initial infection events that had given rise to exactly E_0 infected individuals at $t = 0$.

The delayed terms in the equations for groups *E* and *I* ensure individuals spend exactly T_E and T_I time units in these groups. The outflow at moment t is equal to the inflow at $(t - T_E)$ or $(t - T_I)$ respectively. This is equivalent to both the residence time distribution and outflow rate being equal to the shifted Dirac delta function (unit impulse) of the time spent in the group, or to a rectangular survival function in the group. At the same time, it can be viewed as an extreme case of the Erlang (or Gamma) distribution of residence times as in groups *B* and *C*, but if the number of dummy variables (N_{gB} and N_{gC}) went to infinity.

The initial number of immune schoolboys (i.e., those in group R at t=0, or $R_{t=0}$) was assumed to be non-zero for two reasons:

1) Some schoolboys might have been immune to influenza A/H1N1. The strain that caused the "children's pandemic" had not been widespread for 15-20 years before the event, but the

original paper [5] reported that 630 boys were vaccinated with Fluvirin (that had no H1N1 component) in October 1977. Thus, some cross-immunity might have played a role [21, 22].

2) Non-zero $R_{t=0}$ can be a crude way of accounting for contact network effects when using a globally mixed compartmental model: the real contact network is usually far from globally connected. The original publication [5] reported that the schoolboys were of very different ages, 10 to 18 y.o., and “113 boys of 10-13 y.o. were in the junior house, and the rest were divided into 10 houses of about 60 boys each”. Thus, some boys could have been shielded from the infection by saturation of their local contact networks, thus making them unavailable for infection (hence, immune in the model), while still being biologically susceptible.

Table 1. Variables, parameters, and fitting results of the SEIBCR model (1). Parameters that were estimated in the fitting process are marked with “^{e}”.

Symbol	Description	Dimension	Min. value	Max. value	Optimal value
$S(t)$	Number of susceptibles at time t	pers.	-	-	Variable
$E(t)$	Number of exposed/infected boys at time t	pers.	-	-	Variable
$I(t)$	Number of infectious boys at time t	pers.	-	-	Variable
$B_j(t)$	Number of “confined to bed” boys at time t; j is the index for dummy cascade variables, $B(t) = \sum_j B_j(t)$	pers.	-	-	Variable
$C_j(t)$	Number of convalescent boys at time t; j is the index for dummy cascade variables, $C(t) = \sum_j C_j(t)$	pers.	-	-	Variable
$R(t)$	Number of recovered/immune boys on day t	pers.	-	-	Variable
β_I	Infectivity coefficient for “*I*” state	1/day	0	30	4.3757^{e}
T_E	Residence time in “*E*” group	day	0.1	5	2.984^{e}
p_{noE}	Probability for an infected person to have no “*E*” state and go to “*I*” state immediately	-	0%	100%	67.87%^{e}
$E_{t=0}$	Initial number of *E* individuals at $t = 0$	pers.	0.01	5	0.0429^{e}
T_I	Residence time in “*I*” group	day	0.1	5	1.860^{e}
δ	Average “*B*”→“*C*” progression rate	1/day	0.01	1.0	0.3341^{e}
N_{gB}	The number of dummy groups in “*B*”	-	10	10	10
ε	Average “*C*”→“*R*” progression rate	1/day	0.01	1.0	0.5253^{e}
N_{gC}	The number of dummy groups in “*C*”	-	10	10	10
N	Population size	pers.	763	763	763
$R_{t=0}$	Initial number of immune persons at $t = 0$	pers.	0	251-$E_{t=0}$	248.74^{e}
Δt	Time-shift for data (the first day in the original data corresponds to day Δt, and the model’s initial state is at day $t = 0$)	day	3	3	3
$\mathcal{R}_0$	Basic reproduction number	-	-	-	8.14
AR	Attack rate	pers.	-	-	512.07
$RMSE_B$	Root-mean-square error in *B*	pers.	-	-	3.3543
R_B^2	Coefficient of determination in *B*	-	-	-	99.8783%
$RMSE_C$	Root-mean-square error in *C*	pers.	-	-	7.0878
R_C^2	Coefficient of determination in *C*	-	-	-	98.6743%

Fitting to data

The model was fitted to data by minimising the weighted sum of squared residuals (f) for *B*, *C*, and the attack rate (note that the model attack rate is calculated as $S(t=0)-S(t=\mathrm{end})+E(t=0)$, and r_{AR}^2 is the squared residual in attack rate):

$$
\begin{aligned}
f &= [r_B^2 + w_C r_C^2 + w_{AR} r_{AR}^2] \rightarrow \min \\
r_B^2 &= \sum_t \left(\left(\sum_{j=1}^{N_{gB}} B_j(t) \right) - B^{data}(t) \right)^2 \\
r_C^2 &= \sum_t \left(\left(\sum_{j=1}^{N_{gC}} C_j(t) \right) - C^{data}(t) \right)^2 \\
r_{AR}^2 &= (S(t=0) - S(t=\mathrm{end}) + E(t=0) - AR^{data})^2
\end{aligned}
\qquad (2)
$$

$B^{data}(t)$ and $C^{data}(t)$ are the curves shown in Fig.1, but shifted Δt days to the right in order to have the first cases of infection at $t=0$ exactly Δt days before the first *B*-cases in the data. $AR^{data} = 512$ pers. is the target attack rate. Weights in the target function (2) were heuristically chosen to be $w_C = 0.4$ and $w_{AR} = 10$. (See the Appendix Section 4 for alternative variants and discussion, including $w_{AR} = 0$).

We used a library BFGS gradient descent method for numerical optimisation. To alleviate the problem of local minima, each fit was repeated 50 times from random initial points within the permitted range of parameters and the solution with the best fit to the data was chosen.

Sensitivity and variability analyses

Three sensitivity analyses were performed:

1. An analysis of variability in the model-generation process: we studied how stable the features predicted by the flexible discrete-time model were when a small random noise was applied to the boarding school data. In other words, it was an analysis of the stability of the final model's structure.
2. An analysis of the “plausible set” of the parameters of the final DDE model, i.e., the range of parameters that do not contradict the observed data too much (equivalently, the goodness-of-fit of the model with these parameters is within reasonable limits). To achieve this, we created a lattice over the initial number of immune individuals ($R_{t=0}$) and the infectivity coefficient (β_I) and re-optimised all other “free” parameters at each node of the lattice with $R_{t=0}$ and β_I fixed at their lattice values.
3. A Bayesian analysis of a simplified discrete-time model analogous in structure to our final DDE model. This analysis was carried out with the Stan statistical package [23, 24]. The goal was to determine the limits of identifiability of the model's parameters.

Simplified discrete-time model

The simplified discrete-time model (eq. (3) below) was built for and used in the Bayesian analysis.

Model (3) operated on 1-day time-steps, but was similar to the DDE model (1) in the main features: groups *S*, *E*, *I*, *B*, *C*, and *R*, Dirac-δ-distributed residence times in groups *E* and *I* ($T_E = 3$ days, $T_I = 2$ days; implemented with dummy “memory” subgroups E_1, E_2, E_3, I_1, I_2 that “remember” the preceding values of E and I), and Erlang-like residence times in groups *B* and *C* (implemented with

$N_{gB} = N_{gC} = 2$ consecutive dummy subgroups in each group). In the type of equations, the simplified model was similar to the flexible discrete-time model (A2) (Appendix, Section 3). The parameters are described in Table 2. The model equations are as follows:

$$
\begin{aligned}
& infected(t) = S(t)\left(1 - \left(1 - \frac{1}{N}\right)^{\beta_I\left(I_1(t)+I_2(t)\right)}\right) \\
& S(t+1) = S(t) - infected(t) \\
& E_1(t+1) = (1 - p_{noE})infected(t) \\
& E_2(t+1) = E_1(t) \\
& E_3(t+1) = E_2(t) \\
& I_1(t+1) = p_{noE}infected(t) + E_3(t) \\
& I_2(t+1) = I_1(t) \\
& B_1(t+1) = B_1(t) + I_2(t) - \delta^* B_1(t) \\
& B_a(t+1) = B_a(t) + \delta^*\left(B_{a-1}(t) - B_a(t)\right), \;\; a = 2, \dots, N_{gB} \\
& C_1(t+1) = C_1(t) + \delta^* B_{N_{gB}}(t) - \varepsilon^* C_1(t) \\
& C_a(t+1) = C_a(t) + \varepsilon^*\left(C_{a-1}(t) - C_a(t)\right), \;\; a = 2, \dots, N_{gC} \\
& R(t+1) = R(t) + \varepsilon^* C_{N_{gC}}(t) \\
& S(t=0) = S_{t=0} \\
& E_1(t=0) = (1 - p_{noE})E_{t=0} \\
& E_2(t=0) = E_3(t=0) = 0 \\
& I_1(t=0) = p_{noE}E_{t=0} \\
& I_2(t=0) = 0 \\
& B_a(t=0) = 0, \;\; a = 1, \dots, N_{gB} \\
& C_a(t=0) = 0, \;\; a = 1, \dots, N_{gC} \\
& R(t=0) = N - S_{t=0} - E_{t=0}
\end{aligned}
\qquad (3)
$$

Numerical solutions of the discrete-time model (3) (just as of the flexible model (A2)) can be calculated much faster than solutions of ODE or DDE models with adaptive time-lattices. Model (3) may be easier to interpret than the DDE model (1).

Results

Flexible discrete-time model

The full results of the flexible discrete-time model's fitting are presented in the Appendix, Section 3. The estimates so obtained gave rise to the final form of the DDE model (1). The goodness-of-fit of the flexible model was expectedly very high ($RMSE_B = 3.3$ pers., $RMSE_C = 6.8$ pers., $AR = 511.9$ pers.), but it was an overfitting due to the high number of free parameters.

Basic fitting of the DDE model

The results of fitting the DDE model (1) are shown in Table 1 and Figure 2. The fit was extremely good in the *B*-curve ($RMSE_B = 3.4$ pers.), quite reasonable in the *C*-curve ($RMSE_C = 7.1$ pers.), and the attack rate was reproduced very well ($AR = 512.07$ pers., $AR^{data} = 512$ pers.).

The underlying "epidemiologic mechanism" was not the "typical" one: the infection was estimated to be highly contagious ($\mathcal{R}_0 = 8.14$), the initial number of immune schoolboys was high ($R_{t=0} \approx 249$ pers.), and the outbreak was stopped by a total exhaustion of all susceptibles.

The "typical" mechanism followed in previous studies would imply a lower $\mathcal{R}_0$ (typically, $\mathcal{R}_0$=1–4 for pandemic influenza in general populations [2, 25]), zero or very low $R_{t=0}$, and the outbreak would be stopped by the effective reproduction number dropping below 1, with some susceptibles still remaining.

The mean generation time was $GT = (1 - p_{noE})T_E + T_I/2 = 1.888$ days.

The unusual profile in the *I* curve (Fig.2, *I*-pane) was created by most infections taking place almost simultaneously and then, due to some infectees bypassing the *E* group, arriving in the *I* state in two bursts – one immediate, and one with the T_E lag.

Additional variants of model (1)'s fittings are shown in the Appendix, Section 4. These include scenarios without AR targeting ($w_{AR} = 0$) and without the "E-bypass" mechanism ($p_{noE} = 0$).

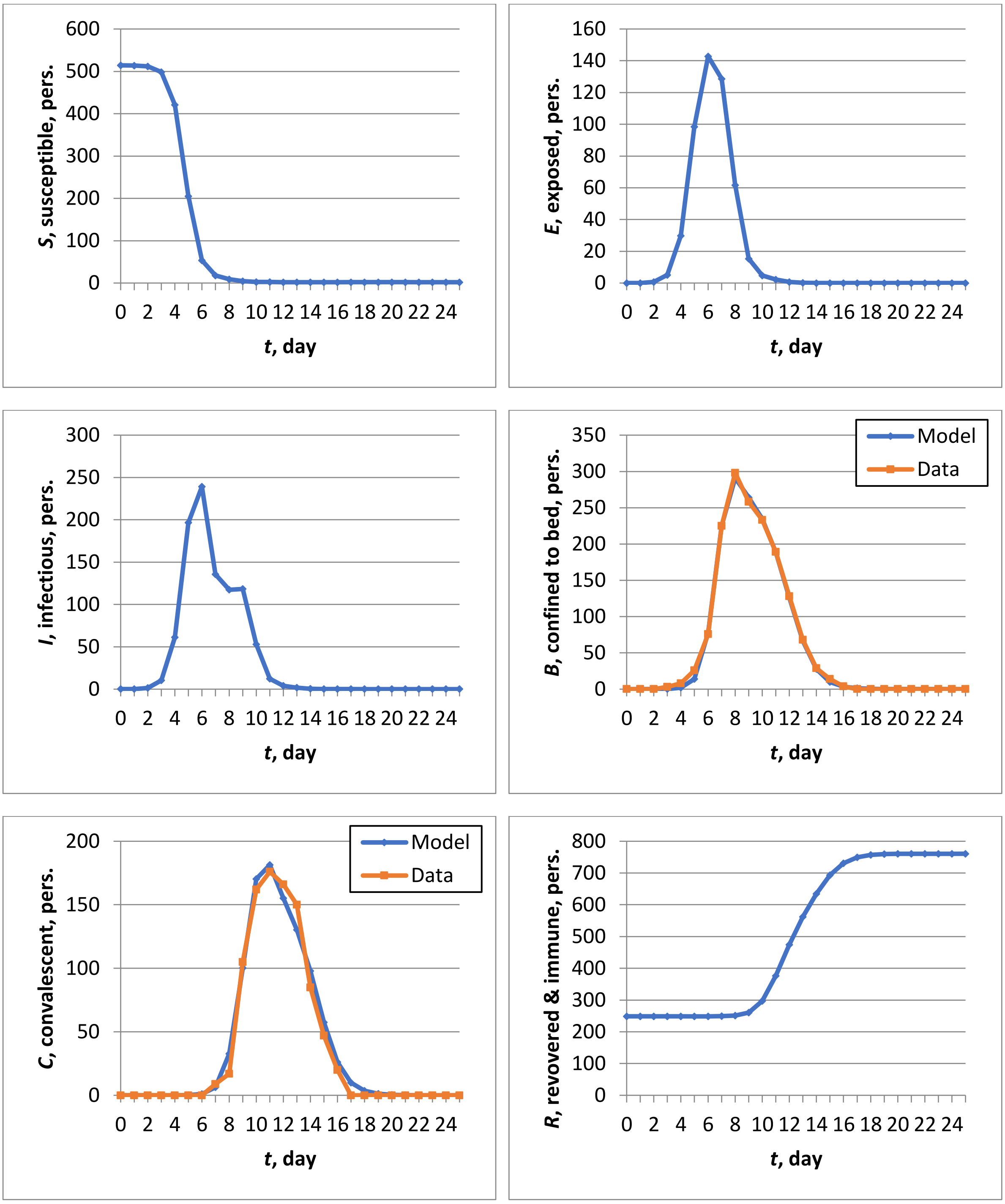

Figure 2. DDE model (1) variables and fit to the data in variables B and C (in brown) with the initial number of immune individuals ($R_{t=0}$) being esimated. Variables B and C are sums of the relevant dummy variables B_j and C_j respectively.

Sensitivity

The analysis of the stability of features predicted by the flexible model is presented in the Appendix, Section 5. All the main qualitative features of the final continuous model (1) remained unchanged except for the shape of the distribution of residence time in group *C*.

The estimation of the "plausible set" of parameters of the DDE model (1) is presented in the Appendix, Section 6. The best-fitting variant was with a high $R_{t=0}$ value and moderate $\mathcal{R}_0$, but there were alternative variants with no initial immunity ($R_{t=0} = 0$) and with a still reasonable fit to the data.

The Bayesian fitting and analysis of the simplified discrete-time model (3) were carried out with the Stan package [23, 24]. We chose to use model (3) rather than the DDE model (1) to avoid the use of DDE solvers (which are absent in Stan). The key features of model (1) were preserved in model (3).

The simplified model (3) was fitted to the boarding school data using the Bayesian procedure. There were three fitting scenarios:

1. with fixed $S_{t=0} = 512$ pers. (which is equivalent to fixing $R_{t=0}$ in other our models),
2. with estimated $S_{t=0}$ (hence, $R_{t=0}$ is effectively estimated too),
3. with fixed $S_{t=0} = N - E_{t=0}$ (which means $R_{t=0} = 0$, $S_{t=0} \approx N$).

The prior distributions of parameters were relatively non-informative:

$$\beta_I \sim \text{Normal}(\mu = 5, \sigma^2 = 2^2) \in [0, 30];$$

$$p_{noE} \sim \text{Uniform}([0, 1]);$$

$$\delta^* \sim \text{Uniform}([0.01, 1]);$$

$$\varepsilon^* \sim \text{Uniform}([0.01, 1]);$$

$$E_{t=0} \sim \text{Normal}(\mu = 1, \sigma^2 = 1) \in [0.01, 5];$$

$$S_{t=0} \sim \text{Uniform}([0, 763]).$$

The model's likelihood function was a product of two likelihood functions (for *B* and *C*) based on two statistical assumptions:

$$B^{data}(t) \sim \text{NegativeBinomial}\left(\mu_{B(t)}, \Phi_B\right),$$
$$\mu_{B(t)} = B^{model}(t) = \mathbb{E}\left(B^{data}(t)\right),$$
$$\text{Var}\left(B^{data}(t)\right) = \mu_{B(t)} + \frac{\left(\mu_{B(t)}\right)^2}{\Phi_B},$$
$$\Phi_B \sim \text{Normal}(\mu = 0, \sigma^2 = 5^2) \in [0, \infty)$$

and

$$C^{data}(t) \sim \text{NegativeBinomial}\left(\mu_{C(t)}, \Phi_C\right),$$
$$\mu_{C(t)} = C^{model}(t) = \mathbb{E}\left(C^{data}(t)\right),$$
$$\text{Var}\left(C^{data}(t)\right) = \mu_{C(t)} + \frac{\left(\mu_{C(t)}\right)^2}{\Phi_C},$$
$$\Phi_C \sim \text{Normal}(\mu = 0, \sigma^2 = 5^2) \in [0, \infty).$$

In this model, the attack rate data was not directly targeted.

Table 2. Variables, parameters, and fitting results of the simplified discrete-time SEIBCR model (3). Parameters that were estimated in the fitting process are marked with “^{e}”. Values with posterior distributions are reported as “mean (95% Bayesian Credible Interval)”.

Symb.	Description	Dim.	Scenario 1, fixed $S_{t=0}$	Scenario 2, estimated $S_{t=0}$	Scenario 3, fixed $S_{t=0} = N - E_{t=0}$
$S(t)$	Number of susceptibles on day t	pers.	Variable	variable	variable
$E_i(t)$	Number of exposed/infected boys on their i-th day of infection on day t	pers.	Variable	variable	variable
$I_i(t)$	Number of infectious boys on their i-th day of the state on day t	pers.	Variable	variable	variable
$B_a(t)$	Number of “confined to bed” boys on day t, a-th dummy subgroup	pers.	Variable	variable	variable
$C_a(t)$	Number of convalescent boys on day t, a-th dummy subgroup	pers.	Variable	variable	variable
$R(t)$	Number of recovered/immune boys on calendar day t	pers.	Variable	variable	variable
$S_{t=0}$	Initial number of susceptible persons at $t = 0$	pers.	512	573.26(453.18-718.66) ^{e}	762.19(761.16 -762.75) ^{e}
β_I	Infectivity coefficient for “I” state	1/day	6.72 (4.54-9.29) ^{e}	6.20 (3.85-9.05) ^{e}	4.82 (2.93-7.72) ^{e}
p_{noE}	Probability for an infected person to bypass the “E” state	-	0.66 (0.49-0.85) ^{e}	0.65 (0.47-0.85) ^{e}	0.64 (0.41-0.86) ^{e}
$E_{t=0}$	Initial number of infected at $t = 0$	pers.	0.70 (0.27-1.51) ^{e}	0.72 (0.27-1.56) ^{e}	0.81 (0.25-1.84) ^{e}
δ^*	Progression rate for B-group	1/day	0.75 (0.67-0.83) ^{e}	0.76 (0.67-0.86) ^{e}	0.82 (0.73-0.94) ^{e}
ε^*	Progression rate for C-group	1/day	0.97 (0.88-0.99) ^{e}	0.97 (0.88-1.00) ^{e}	0.96 (0.87-1.00) ^{e}
N	Population size	pers.	763	763	763
Δt	Time-shift for the data	day	3	3	3
AR	Attack rate	pers.	512.5 (511.6-513.4)	573.8 (453.5-719.0)	762.7 (760.8-763.0)
$RMSE_{\tilde{B}}$	Root-mean-square error in *B* of the median solution	pers.	14.366	8.101	25.756
$R^2_{\tilde{B}}$	Coefficient of determination in *B* of the median solution	-	97.68%	99.26%	92.54%
$RMSE_{\tilde{C}}$	Root-mean-square error in *C* of the median solution	pers.	10.160	16.141	44.925
$R^2_{\tilde{C}}$	Coefficient of determination in *C* of the median solution	-	97.17%	92.85%	44.61%

We confirmed the posterior calibration of the simplified model's posterior distributions using Talts et al.'s simulation-based calibration approach [26] (Appendix, Section 7). Therefore, since the 95% CIs are expected to contain the true parameter 95% of the time, the widths of 95% CIs are reliable for drawing inferences on the corresponding parameters.

The results of fitting of the simplified model (3) are shown in Table 2 and Figure AF20 (Appendix, Section 7).

With model (3)'s median solutions, the goodness-of-fit measures (RMSE and R^2) were worse than those for the DDE model (1), but still acceptable in scenarios 1 and 2 (Scenario 1: $RMSE_B = 14.4$ pers., $RMSE_C = 10.2$ pers., $AR = 512.5$ pers.; Scenario 2: $RMSE_B = 8.1$ pers., $RMSE_C = 16.1$ pers., $AR = 573.8$ pers.). Scenario 3 (with, effectively, $R_{t=0} = 0$) did not yield an acceptable fit to the data ($RMSE_B = 25.8$ pers., $RMSE_C = 44.9$ pers., $AR = 762.7$ pers.).

Although the attack rate (AR) was not directly fitted in this model, it was reproduced well in scenario 1 (by indirectly controlling it by setting $S_{t=0} \approx AR$). In scenario 2, AR was still close to the real one – just because of the structure of the model. In scenario 3, AR was unacceptably high (nearly 100%).

The estimated 95% CIs of the parameters (Table 2) showed that none of the parameters of the simplified model (3) could be identified with high precision, yet most of them had moderately wide CIs: about $\pm 42\%$ of the mean value in β_I; about $\pm 30\%$ in p_{noE}; about $\pm 92\%$ in $E_{t=0}$; about $\pm 12\%$ in δ^*; about $\pm 6\%$ in ε^*. With such a small dataset (14 data-points), precise identifiability was not expected.

Reproduction number estimation

We used EpiEstim method and R package [27, 28] to estimate the time-dependent effective reproduction number $\mathcal{R}_t$ (note that $\mathcal{R}_t = \mathcal{R}_0 S(t)/N$) for the given outbreak independently of our models (Appendix, Section 8).

The EpiEstim's results confirmed that the basic reproduction number $\mathcal{R}_0$ of about 10 (which was attained in our model's best scenario) is plausible, as opposed to $\mathcal{R}_0 = 1\text{–}4$ estimates for the general population [2, 25].

Discussion

General modelling

The classical 1978 boarding school dataset was extensively used as an example in various textbooks and in R packages to illustrate the beauty of the SEIR equations in application to real data. It is a well-known example to most mathematical epidemiologists. However, previous works, including Kalachev et al. [16], did not include the total number of schoolboys who fell ill (attack rate, $AR^{data} = 512$ pers.) as given in the original document [5]. Most mathematical epidemiologists focused on fitting the *B*-curve, or both the *B*-curve and *C*-curve, as if this key information of AR had not existed. This is unfortunate, as pointed out by Prof. Michael Li [15] who was the first to attempt to fit both the *B*-curve and the AR. To the best of our knowledge, before our study, it was still an open question whether a biologically reasonable model can be fitted to both the *B*-curve and the AR simultaneously.

Our modelling has shown that the 1978 English boarding school data can be fitted quite well (with good fits in both the *B*- and *C*-curves and in the attack rate) if a model with an appropriate structure is used. The key structural features of our final DDE model (1) were:

1. the actual infectious state (*I*) preceded the observed *B*- and *C*-states of the disease,
2. disease-related groups had non-exponential residence times (Dirac-δ-distributed times in *E* and *I*, Erlang-distributed times in *B*- and *C*-states),
3. some individuals bypassed the *E*-state after infection.

All these features look biologically plausible, although the last one is a more unusual one. They were derived from the given dataset, and so there is no guarantee that they work equally well with other influenza A datasets. Still, we can produce hypotheses that support our interpretation of the features as "plausible":

- The ability of many infectious respiratory diseases to start infectivity before the development of major symptoms is well-known, including for influenza [29, 30]. So, the assumption of existence of the infectious stage *I* that precedes the observed (and, hence, symptomatic) stage *B* appears to be natural.

- Non-exponential disease stage residence time distributions are well-known to be more realistic than exponential ones [31]. The latter are used mostly because of their mathematical simplicity.
  Erlang (Gamma) distributions are “strongly preferred on theoretical grounds” [31]. Dirac-δ-distributed residence time (implemented with DDEs) can be viewed as an extreme case of the Erlang distribution.
  The exact shape of the residence time distribution is likely not critical: any unimodal distribution concentrated around its mean and having very low density near zero will likely work as good as the Erlang distribution does.

- The “E-bypass” mechanism (when a newly infected person skips the exposed stage and develops infectiousness instantly) can be biologically interpreted as an extremely short exposed state. Influenza is known to have a very short generation time [32], and there are case-reports of extremely quick infectiousness development (i.e., very short latent period) [33]. So, for a given highly selected population (schoolboys) and some given strain of influenza A/H1N1 that infected them all, there is no reason to deny a possibility of a quite common ($p_{noE} = 67.9\%$) fast development of infectiousness that, in our model, is interpreted as the “E-bypass”. The boys who did not progress to infectiousness quickly, can be hypothesised to have been more resistant or having partial immunity.
  On the other hand, there is still the possibility that the “E-bypass” we have identified is just an artefact of fitting the given dataset, although we think this unlikely. Technically, the “E-bypass” was justified by the survival function in *E* obtained in the flexible model (A2) (Fig. AF5).

Some parameters of our model (1) differed from the usual ones:

- The initial number of immune individuals ($R_{t=0}$) was substantially non-zero (discussed in "Final DDE model"), although we found reasonably well-fitting parameter sets with $R_{t=0} = 0$ (Appendix, Section 6).
- The residence time in the infectious *I*-state was $T_I = 1.86$ days, which is shorter than the typical reported viral shedding period for influenza A/H1N1 (4.5 days in [33]). This could be explained by a possible virus variation and host population very different from the general population. But also, the model *I*-group is "infectious individuals before being detected as influenza cases and confined to bed". Thus, the *I*-state ends not with an actual cessation of infectiousness, but with quarantining; and some *B*-individuals can biologically be still somewhat infectious, but the quarantine measures make them effectively non-infectious for the remaining susceptibles. So, T_I can be shorter than "infectiousness time" observed in practical influenza A/H1N1 studies.

The feature of the 1978 boarding school data is that the outbreak was very quick (14 days) and almost all infections must have taken place within a few days. It made the disease progression "schedules" of all infectees mostly synchronised in time. Thus, the observed *B*- and *C*-curves represent much more the dynamics of progression of the disease (i.e., its natural history), rather than the dynamics of new infections. This is the likely reason why our epidemiologic models (either with an extremely flexible description of disease stage progression, or with a custom-tailored disease progression part) were able to fit the data so well.

The classic SEIR models with exponentially-distributed residence times were very successful in explaining large-scale dynamic transitions [34] and were successfully fitted to epidemic curves in large populations [1, 2]. But the classic SIR/SEIR models are ineffective in approximating the 1978 boarding school dataset [15]. Presumably, this is because in slower, non-synchronised, and many-generation processes, only the mean residence times matter. Yet, the school outbreak is different (fast, synchronised, few generations of infections), and the classic SEIR models cannot conform to its features.

Flexible model approach

Initially, we found that standard SEIR-type models gave poor fits to the data (Appendix, Section 2). So, some "deformation" of the models was needed. It was possible to introduce a time-dependent infectivity $\beta_I(t)$, but we considered it as a less plausible assumption, given such a short time-interval. Our aim in this work was to find a biologically reasonable explanation, and we did not attempt to rule other possibilities out.

Thus, we proposed a flexible model approach: we constructed an intentionally overly flexible discrete-time model (A2) that made practically no assumptions about the distribution of the residence times in the model groups (Appendix, Section 3), fitted the flexible model to the data, and, by reviewing the fitted survival functions, determined what distributions should be used in the simpler final model (1). Those distributions of residence times turned out to be Dirac delta distribution (unit impulse) for *E* and *I* (i.e., constant residence time in *E* and *I*), and Erlang (Gamma) for *B* and *C*. Hence, the DDE framework and dummy groups mechanism were used in the final model (1).

To show the wider applicability of the model-generation approach, we used it to build an optimised SEIR-like model for COVID-19 outbreak in Hong Kong in 2022 – see the Appendix, Section 10.

Generality

We expect the model-generation approach (i.e., the “flexible model” approach) to be applicable in many situations when a better-fitting, yet simple *candidate* model is needed. The generated simple model is expected to work well with the given dataset (that was used in the model generation) and to reveal some hypotheses about the underlying biological processes – this is why it is only a *candidate* model. Confirmation or rejection of the hypotheses requires further research and bigger datasets.

Obviously, the candidate model is likely to fit well only datasets of the similar nature (in our case, quick outbreaks in tightly connected small populations), and it cannot replace the classic SEIR model everywhere. The alternative example of the 2022’s COVID-19 outbreak in Hong Kong (Appendix, Section 10) produced an optimised model somewhat different from the boarding school’s one, but still quite close in the general principles: separated-from-zero distribution of residence times in *E*, *E*-bypass, short infectiousness time.

Reproduction numbers

Our “central” estimate of the basic reproduction number for the influenza A/H1N1 boarding school outbreak is $\mathcal{R}_0 = 8.14$, which is much higher than typical estimates of $\mathcal{R}_0 =$ 1–4. The difference can be explained by higher contact rates between schoolchildren than between the average members of the general population and by the boarding school setting increasing the contact rates even more as compared to “average schoolchildren” who are not confined to in-school-only contacts (Appendix, Section 8).

The independent estimation of $\mathcal{R}_0$ with the EpiEstim method [27, 28] supported the possibility of high $\mathcal{R}_0$ ($\mathcal{R}_0 =$ 5–10) – see the Appendix, Section 8.

Goodness of fit comparison

A summary table and plot of goodness-of-fit of all models dealing with the 1978 boarding school data are presented in Section 11 of the Appendix.

The final DDE model (1) approximated the *B*-curve with almost 5 times smaller RMSE (root-mean-square error) than the SIR and SEIR models that reproduced AR (RMSE 3.3543 versus 16.5250 and 16.0462). For the SIR and SEIR models targeting only the *B*-curve and having AR near 100%, the difference was about 3.5 times (RMSE 3.3543 versus 12.3329 and 11.0997).

The flexible discrete-time model (A2) was expectedly better in goodness-of-fit than the DDE model (1), but the difference was under 5% in RMSE ($RMSE_B^{(A2)} = 3.2842$, $RMSE_B^{(1)} = 3.3543$,

$RMSE_C^{(A2)} = 6.7813$, $RMSE_C^{(1)} = 7.0878$). This means that the DDE model (1) is practically as effective in approximating the dataset as the flexible model (A2), while it has only a few free parameters more than the standard SEIR model (with the *C*-group added).

The simplified discrete-time model (3) was considerably worse than the DDE model (1) in goodness-of-fit: in the "forced correct AR" Scenario 1, it was 4.3 times worse in *B* ($RMSE_B^{(3)} = 14.366$) and 1.4 times worse in *C* ($RMSE_C^{(3)} = 10.160$) than model (1); in the "no AR targeting" Scenario 2, there was a significant difference in AR ($AR^{(3)} = 573.8$, $AR^{data} = 512$) and the fits in *B* and *C* were about 2.3 times worse than in model (1) ($RMSE_B^{(3)} = 8.101$, $RMSE_C^{(3)} = 16.141$). This was expected because model (3) was limited by design to fixed integer residence times in *E* and *I*. Nevertheless, even in this limited form, it performed slightly better than SIR/SEIR models.

Model fitting and identifiability

The DDE model (1) fitted the boarding school data well. Yet, it technically depended on the heuristically chosen weights w_C and w_{AR} (eq. (2)). In fact, the exact values of these weights were not critical, and the model produced quite similar fits with different weights.

$w_C = 0.4$ reflected the notion that fitting in the *C*-curve was not as important as the one in the *B*-curve.

w_{AR} and AR-targeting were introduced as one of the main features of our modelling process that, unlike other studies, tried to reproduce the observed attack rate together with the *B*- and *C*-curves. Yet, the numerical experiments showed that much lower values of w_{AR} produced practically the same fits. And even if AR-targeting was switched off ($w_{AR} = 0$), the adapted model (1) produced AR quite close to the observed one (518 vs. 512 pers., see the Appendix, Section 4).

Naturally, there is a question of identifiability. It splits into several parts:

1. The identifiability of the *structure* of the optimised model, i.e., how stable the *qualitative* results of the flexible discrete-time model are. As shown in the Appendix Section 5, this qualitative model-generation process is fairly stable to a reasonable random noise in the observed data.
2. The identifiability of the parameters of the final DDE model (1). This question is, in our opinion, less important because we do not claim our DDE model to be universally applicable. Furthermore, the small size of the 1978 boarding school dataset (14 actual data-points) is not conducive to precise technical identifiability of the parameters.
   Nevertheless, we carried out two analyses regarding the issue:
   a. An estimation of the "plausible set" of parameters – in the sense of "what parameter values permit a still acceptable fit to the data" (Appendix, Section 6). It revealed that the DDE model is capable of reasonable fits to the data in different "modes" having different biological interpretations. In the "optimal" mode, the AR is controlled mostly with the suitable initial number of immune individuals ($R_{t=0} \approx N - AR^{data}$); but there exists a mode with $R_{t=0} = 0$, well-reproduced AR, and goodness-of-fit measures considerably better than those of the standard SIR and SEIR models. This shows the potential of model structure being "custom-tailored" to a given dataset.

b. A formal Bayesian fitting procedure of a simplified discrete-time model (3) that replicated the main features of the DDE model (1) ("Sensitivity" section above and the Appendix, Section 7). The widths of obtained posterior credible intervals were expectedly not small, but still could be named "moderate". As shown by a quantitative calibration analysis in Section 7 of the Appendix, the model's posterior distributions were well-calibrated in scenarios 1 and 2, and so its CIs captured the data without any impact from the priors. This indirectly supported the practical identifiability of the DDE model (1).
   The analysis was not suitable for a pinpoint estimation of the parameters, but showed that the general *modus operandi* of the model remained stable.

Drawbacks

Due to the nature of the "resulting" models (1) and (3), there are numerous "weak points":

1. Biological assumptions:
   1.1. "*E*-bypass": not a typical assumption for SEIR-type models, but see its discussion in "General modelling" above.
   1.2. Significant amount of initially immune individuals ($R_{t=0}$): this feature definitely improves the fit in AR both in SIR/SEIR models and in our models, albeit $R_{t=0} = 0$ is traditionally assumed in outbreaks models. See the discussion of this point in "Final DDE model".
   1.3. Dirac-δ-distributed residence times in groups *E* and *I*: these are a simplifying idealisation. But the same could be said for the "standard" constant progression rates of a SEIR model (that result in the exponential distribution of residence times). Both are one-parameter distributions with an easily controlled mean value. Both have distribution shapes that are "biologically questionable".
   1.4. Semi-arbitrary shapes of the Erlang residence time distribution in *B* and *C* (controlled by the number of dummy stages $N_{gB} = N_{gC} = 10$).
2. The fitting procedure depended on heuristically chosen weights in the target function (2). Despite our analysis had shown that changing the weights within sensible limits did not change the qualitative results, the weights still affect the exact numerical estimations of model parameters.
3. Identifiability: due to a small size of the dataset (14 points), the parameters of either model cannot be precisely identified. Furthermore, our sensitivity analysis for model (1) revealed that it can have satisfactory fits to the data with structurally very different parameters (and, hence, *modi operandi* of the model).
4. Model (3) had pre-fixed residence times in *E* and *I*, and those were not optimised (just chosen as close to the parameters of model (1) as possible).
5. The Bayesian fitting of model (3) depended on the choice of prior distributions – just as any Bayesian fitting. We tried to choose as uninformative priors as possible.
6. Our models are still Markovian with regard to disease stages: the residence time in each stage is independent of the previous or subsequent stages. Likely, it is not so in reality. Furthermore, the original paper [5] mentioned that the time in bed and in the convalescent state depended on disease severity and, hence, were correlated.

7. Our final model (1) may be not unique: there might exist other models that have slightly more parameters than the SEIR model and give a good fit to the dataset.

Conclusion

Our primary objective was to resolve the puzzle of whether a biologically sound model can fit the classical textbook 1978's boarding school influenza A/H1N1 outbreak data. We successfully built such a model – the DDE model (1).

Our model (1) reproduced the observed attack rate very well, while the previous "textbook solutions" usually had 100% or near-100% attack rate. The quality of fit in the *B*- and *C*-curves was also 3-5 times better in our model than in many textbook models.

Our model (1) differed by having non-standard, narrow distributions of residence times in the disease-related groups. Most infections in the outbreak were "synchronised" (occurred within a few days), and this permitted the non-standard disease-progression model to fit the data well. Model (1) was built as a simplification or conceptualisation of an intentionally overly flexible model fitted to the same data. The resulting model might have been a mild overfitting, but it was indirectly supported by plausible results of fits without targeting the attack rate (simulated $AR = 518$, while $AR^{data} = 512$). The final delay differential equation model (1) with multi-stage groups may look complex, but the actual number of free parameters is practically the same as for one-stage models.

The final model (1) might be considered as "biologically plausible" and possibly more realistic – for the given dataset – than the standard SEIR model. The residence time distributions and other biological assumptions were chosen on the basis of fitting the overall outbreak curve and general biological speculations (see "General modelling"). The small size of the dataset did not permit a precise formal identification of model's parameters. The fit to the data depended on heuristically chosen weights.

The overall approach of custom-tuning the structure of a model (especially the distributions of residence times) to a given dataset is likely not new, but it still can be useful for producing "hypothetical models" or candidate models in other studies.

We see the main merit of our study and its final DDE model (1) not in the estimated epidemiological parameters, but in the overall model structure (including residence time parameters) and insights that it has provided.

Funding

The work was partially supported by the Hong Kong Research Grants Council Theme-based Research Scheme fund (T24-508/22-N), Collaborative Research Grant (C5079-21G), the National Science Foundation of China (Grant No. 12250410241), and Australian Research Council (DP240102585).

Contributions

KKA and DH conceived the work. KKA – programming, calculations, drafting the paper. LL and HD – programming and calculations. DH and QL secured the funding. DH supervised the whole project. LS, HD, and QL reviewed the work critically. All authors approved the submission.

Ethics

This work did not require an ethical approval because it dealt with historical depersonalised data only.

Artificial Intelligence statement

No artificial intelligence (AI) systems were used in this research project or writing the paper.

Data accessibility

The outbreak data-curves are available in the R package "outbreaks" [20] and in JHU’s COVID-19 Tracker Repository. The Matlab code to reproduce the final model of this paper and RStan code to reproduce the simplified discrete-time model are available in the Zotero repository: https://www.zotero.org/groups/5523146/materials_avilov_et_al_boardingschool_jrsinterface2024.

Conflict of interest declaration

The authors declare that they have no competing interests.

Acknowledgements

The authors thank Prof. Michael Yi Li (University of Alberta) for pointing out this topic to us.

REFERENCES

[1] Bootsma, M.C. & Ferguson, N.M. 2007 The effect of public health measures on the 1918 influenza pandemic in US cities. *Proceedings of the National Academy of Sciences* **104**, 7588-7593.
[2] He, D., Dushoff, J., Day, T., Ma, J. & Earn, D.J. 2013 Inferring the causes of the three waves of the 1918 influenza pandemic in England and Wales. *Proceedings of the Royal Society B: Biological Sciences* **280**, 20131345.
[3] Camacho, A., Ballesteros, S., Graham, A.L., Carrat, F., Ratmann, O. & Cazelles, B. 2011 Explaining rapid reinfections in multiple-wave influenza outbreaks: Tristan da Cunha 1971 epidemic as a case study. *Proceedings of the Royal Society B: Biological Sciences* **278**, 3635-3643.
[4] Mizumoto, K. & Chowell, G. 2020 Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020. *Infectious disease modelling* **5**, 264-270.
[5] Anonymous. 1978 Influenza in a boarding school. *British Medical Journal* **1**, 578.
[6] Francis, M.E., King, M.L. & Kelvin, A.A. 2019 Back to the future for influenza preimmunity—Looking back at influenza virus history to infer the outcome of future infections. *Viruses* **11**, 122.
[7] Davies, J., Smith, A., Grilli, E. & Hoskins, T. 1982 Christ's Hospital 1978–79: An account of two outbreaks of influenza A H1N1. *Journal of Infection* **5**, 151-156.
[8] De Vries, G., Hillen, T., Lewis, M., Müller, J. & Schönfisch, B. 2006 *A course in mathematical biology: quantitative modeling with mathematical and computational methods*, SIAM.
[9] Grinsztajn, L., Semenova, E., Margossian, C.C. & Riou, J. 2021 Bayesian workflow for disease transmission modeling in Stan. *Statistics in medicine* **40**, 6209-6234.
[10] Martcheva, M. 2015 *An introduction to mathematical epidemiology*, Springer.
[11] Raissi, M., Ramezani, N. & Seshaiyer, P. 2019 On parameter estimation approaches for predicting disease transmission through optimization, deep learning and statistical inference methods. *Letters in biomathematics* **6**, 1-26.
[12] Keeling, M.J. & Rohani, P. 2011 *Modeling infectious diseases in humans and animals*, Princeton university press.
[13] Hao, Y.J., Wang, Y.L., Wang, M.Y., Zhou, L., Shi, J.Y., Cao, J.M. & Wang, D.P. 2022 The origins of COVID-19 pandemic: A brief overview. *Transboundary and Emerging Diseases* **69**, 3181-3197.
[14] Rozo, M. & Gronvall, G.K. 2015 The reemergent 1977 H1N1 strain and the gain-of-function debate. *MBio* **6**, 10.1128/mbio. 01013-01015.
[15] LI, M.Y. 2023 An Epidemic Enigma: Challenges in Modeling the Influenza Epidemic in a Boarding School. *2023 Canadian Mathematical Society Summer Meeting, Ottawa, Canada, June 2-5th, 2023*.
[16] Kalachev, L., Landguth, E.L. & Graham, J. 2023 Revisiting classical SIR modelling in light of the COVID-19 pandemic. *Infectious Disease Modelling* **8**, 72-83.
[17] Butler, D. 2014 Models overestimate Ebola cases. *Nature* **515**, 18.
[18] Baraniuk, C. 2020 What the Diamond Princess taught the world about covid-19. *Bmj* **369**.
[19] Diekmann, O., Heesterbeek, J.A.P. & Metz, J.A. 1990 On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. *Journal of mathematical biology* **28**, 365-382.
[20] Jombart, T., Frost, S., Nouvellet, P., Campbell, F. & Sudre, B. 2017 outbreaks: A Collection of Disease Outbreak Data. *R package version* **1**.
[21] Gatti, L., Koenen, M.H., Zhang, J.D., Anisimova, M., Verhagen, L.M., Schutten, M., Osterhaus, A. & van der Vries, E. 2022 Cross-reactive immunity potentially drives global oscillation and opposed alternation patterns of seasonal influenza A viruses. *Scientific reports* **12**, 8883.
[22] Hillaire, M.L., van Trierum, S.E., Kreijtz, J.H., Bodewes, R., Geelhoed-Mieras, M.M., Nieuwkoop, N.J., Fouchier, R.A., Kuiken, T., Osterhaus, A.D.E. & Rimmelzwaan, G.F. 2011 Cross-protective

immunity against influenza pH1N1 2009 viruses induced by seasonal influenza A (H3N2) virus is mediated by virus-specific T-cells. *Journal of General Virology* **92**, 2339-2349.
[23] Guo, J., Gabry, J., Goodrich, B. & Weber, S. 2020 R Package 'rstan'. *https://cran.r-project.org/web/packages/rstan/index.html*.
[24] Houston, R. 2016 An Introduction to Stan and RStan.
[25] Biggerstaff, M., Cauchemez, S., Reed, C., Gambhir, M. & Finelli, L. 2014 Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC infectious diseases* **14**, 1-20.
[26] Talts, S., Betancourt, M., Simpson, D., Vehtari, A. & Gelman, A. 2018 Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.
[27] Cori, A., Cauchemez, S., Ferguson, N.M., Fraser, C., Dahlqwist, E., Demarsh, P.A., Jombart, T., Kamvar, Z.N., Lessler, J. & Li, S. 2020 Package 'EpiEstim'. *CRAN: Vienna Austria*.
[28] Thompson, R.N., Stockwin, J.E., van Gaalen, R.D., Polonsky, J.A., Kamvar, Z.N., Demarsh, P.A., Dahlqwist, E., Li, S., Miguel, E. & Jombart, T. 2019 Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics* **29**, 100356.
[29] Ng, S., Lopez, R., Kuan, G., Gresh, L., Balmaseda, A., Harris, E. & Gordon, A. 2016 The timeline of influenza virus shedding in children and adults in a household transmission study of influenza in Managua, Nicaragua. *The Pediatric infectious disease journal* **35**, 583-586.
[30] Aoki, F.Y. & Boivin, G. 2009 Influenza virus shedding—excretion patterns and effects of antiviral treatment. *Journal of clinical virology* **44**, 255-261.
[31] Krylova, O. & Earn, D.J. 2013 Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of The Royal Society Interface* **10**, 20130098.
[32] Cowling, B.J., Fang, V.J., Riley, S., Peiris, J.S.M. & Leung, G.M. 2009 Estimation of the serial interval of influenza. *Epidemiology* **20**, 344-347.
[33] Carrat, F., Vergu, E., Ferguson, N.M., Lemaitre, M., Cauchemez, S., Leach, S. & Valleron, A.-J. 2008 Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *American journal of epidemiology* **167**, 775-785.
[34] Ferrante, L., Duczmal, L.H., Steinmetz, W.A., Almeida, A.C.L., Leão, J., Vassão, R.C., Tupinambás, U. & Fearnside, P.M. 2021 Brazil's COVID-19 epicenter in Manaus: how much of the population has already been exposed and are vulnerable to SARS-CoV-2? *Journal of Racial and Ethnic Health Disparities*, 1-7.