Proceedings of the Operational Research Society Simulation Workshop 2021 (SW21)
M. Fakhimi, D. Robertson, and T. Boness, eds.

MODELLING PRESYMPTOMATIC INFECTIOUSNESS IN COVID-19

Dr. Russell Cheng
University of Southampton
Highfield, SO17 5BJ
United Kingdom
cheng@btinternet.com

Dr. Christopher Dye
University of Oxford
Mansfield Road, OX1 3SZ,
United Kingdom
christopher.dye@biology.ox.ac.uk

Dr. John Dagpunar
University of Southampton
Highfield, SO17 5BJ
United Kingdom
j.dagpunar@soton.ac.uk

Dr. Brian Williams
SACEMA
Stellenbosch University
South Africa
williamsbg@me.com

ABSTRACT

This paper considers SEPIR, an extension of the well-known SEIR continuous simulation compartment model. Both models can be fitted to real data as they include parameters that can be estimated from the data. SEPIR deploys an additional presymptomatic infectious compartment, not modelled in SEIR but known to exist in COVID-19. This stage can also be fitted to data. We focus on how to fit SEPIR to a first wave of Covid. Both SEIR and SEPIR and the existing SEIR models assume a homogeneous mixing population with parameters fixed. Moreover, neither includes dynamically varying control strategies deployed against the virus. If either model is to represent more than just a single wave of the epidemic, then the parameters of the model would have to be time dependent. In view of this we also show how reproduction numbers can be calculated to investigate the long term overall outcome of an epidemic.

Keywords: Differential equation epidemic models, Parametric models, Effective Reproduction Number, Asymptomatic transmission

1 INTRODUCTION

A parametric SEIR model has been used by the authors in Dye *et al*. (2020) to compare the first wave of the COVID-19 epidemics in different European countries, fitting the model to data using the method of maximum likelihood rather than the Bayesian Markov-chain. The SEIR model is standard but does not include a specific compartment to represent the presymptomatic infectiousness stage which is known to occur in those infected by COVID-19. Presymptomatic infectiousness is a key characteristic of COVID-19 epidemics and contributes strongly to the spread of the disease. He *et al*. (2020), (summarized in Section 2) show that measures for controlling COVID-19 epidemics need to take this into account.

The distinction between ‘presymptomatic’ and ‘asymptomatic’ infections is important. Following the World Health Organization ‘asymptomatic’ people are infected but never develop symptoms while ‘presymptomatic’ people are infected, have not yet developed symptoms but go on to develop symptoms later. Between one third and one half of infected people are asymptomatic. The purpose of this paper is to add a presymptomatic compartment, P, to the SEIR model to estimate the biological and transmission characteristics of presymptomatics explicitly. The SEPIR model, like the SEIR model is a parametric model, and is completely specified once the values of the parameters are all known. Some parameters may be known from other studies, but all the parameters can be estimated by fitting the model to data.

We describe the Maximum-Likelihood (ML) method which produces point estimates of the parameter values (See Cheng, 2017, for example.) In Subsection 4.2 we give a numerical example

fitting the first wave of the COVID-19 epidemic in Switzerland. In particular we show how to calculate, the proportion, P_{pi}, of presymptomatic infections.

In Section 5 we use *bootstrapping* to calculate confidence intervals (CI) for the parameter estimates and calculate confidence limits for estimated trajectories. We have used maximum-likelihood estimation which focuses on point estimation of unknown parameter values. These are all well used methods because they are known for their reliability. We have not used Bayesian estimation which focuses on the posterior distribution of parameters, preferring likelihood methods where the focus is on point estimation. For simplicity we have not investigated both, as this would only be repeating essentially the same conclusions. Other approaches could be used, such as agent-based modelling, but such methods which track individuals, are computationally more demanding. We focus on the ODE approach which is well-known and well-regarded, and leave to others to investigate other approaches.

We focus in this paper on 'first wave' behaviour of the SEPIR model, where this assumption is not unreasonable, once the initial control strategy is in place. We assume that parameter properties are stationary, leaving for elsewhere discussion of situations where time-varying parameters might be used. In the long-term strategies have been deployed to contain a rapidly changing epidemic, like that produced by COVID-19, with model parameters that are time dependent in order to reproduce the trajectory of the epidemic. We consider the epidemic in Switzerland and show how the estimated trajectories can be used for prediction purposes. Compartment models can be used to fit more than one wave, see for example Dagpunar (2020).

In Section 6 we discuss the use of vaccination to control an epidemic. In this case progress of the epidemic is summarised by the *effective* reproduction number R_t, a dynamically varying version of R_0, the (basic) reproduction number. R_0 is generally defined in terms of an idealized epidemic in a homogeneously mixed population, with no control measures, and assuming that every member of the population is susceptible. This assumes that only an initial variant is present. But R_0 can change over time when a new variant emerges as happened in November 2020 in UK with the emergence of the 'Kent variant' b1.117. R_t, on the other hand, can change over time through non-pharmaceutical interventions (NPIs) and changes in the susceptible proportion through infection, vaccination, or waning of immunity from prior infection/vaccination.

When monitoring the overall progress of an epidemic, R_t is more useful than R_0 and, in lay terms, is often referred to as the 'reproduction number'. It can still be defined as the expected number of persons that one infectious person goes on to infect but its value is time dependent because its definition allows for the management of the epidemic, through NPIs and vaccinations, and depends on the proportion of the population that is susceptible, both of which change with time. We show how to calculate R_t and R_0 using the fitted SEPIR model, and discuss how, without vaccination, the final outcome of the epidemic depends on R_0, a celebrated result given by Kermack and KcKendrick (1927) for the SIR model, but which applies also to the SEIR and SEPIR models. We show how the final outcome can only be changed by herd immunity and how this can be accelerated by vaccination.

2 PRESYMPTOMATIC INFECTIONS

In this Section we justify inclusion of the presymptomatic compartment P in the SEPIR model by summarizing the findings of He *et al.* (2020) who investigated the case histories of 77 infector-infectee pairs in each of which an infectious person, the infector, goes on to infect a susceptible person, the infectee. Assuming a mean incubation period as 5.2 days, He *et al*. (2020) estimated the mean *serial interval*, the duration between the start of an infector and an infectee infection, to be 5.8 days. They infer that infectiousness starts 2.3 days after the onset of infection of the infector and peaks just 0.7 days before symptom onset, giving an estimated proportion of infectee infections occurring *before* the onset of infector symptoms as 44%. Infectiousness then declines away over 7 days.

Figure 1 is a schematic diagram showing the infector-infectee relationship as found by He. The arrowed durations show population averages and do not indicate how many individuals have short or possibly even no symptomatic duration. However, we have not included this, given that the present model produces a reasonable estimate of the probability of presymptomatic infection. The red

horizontal arrows depict typical mean durations of the different stages of the infection experienced by the infector. The green horizontal line corresponds to the serial interval. Mean durations are depicted, but all are random variables, so will vary between different infector-infectee pairs. The black dashed curve in Figure 3 depicts the probability density function (PDF) of the serial interval.

The time when the secondary case is infected has to be where the PDF is positive. The shaded area indicates that infection can occur when the infector is presymptomatic with probability indicated by the height of the PDF at that time. The shaded area in the Figure gives the proportion of individuals infected presymptomatically which, in the sample of He *et al*. (2020), was 44%. Higher presymptomatic infection proportions of 48% in Singapore and 62% in Tianjin been given by Tapiwa *et al*. (2020). The practical consequences of these findings are evident, with highly reliable track and tracing required to identify presymptomatic infections.

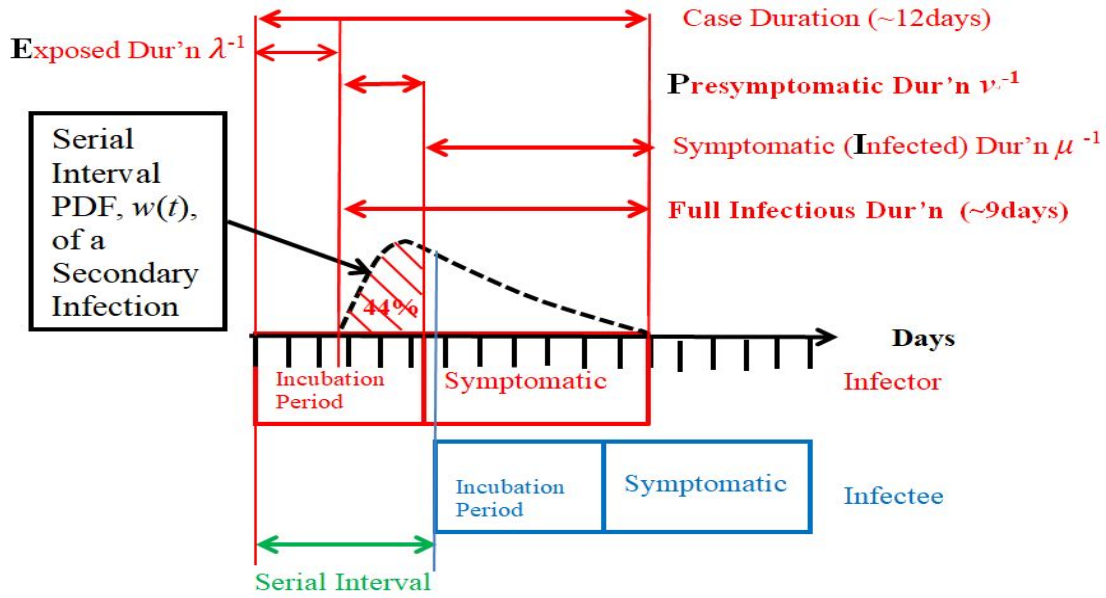

Figure 1: Infector-Infectee Relationship as described by He *et al* (2020). During time **E** the infector is infected but not infectious. During time **P** the infector is asymptomatic but increasingly infectious. During time **I** the infector is symptomatic but their infectiousness falls over time.

3 THE SEPIR MODEL

Our model is similar to the SEIR model described in the Supplementary Materials of Dye *et al*. (2020) but augmented with an extra compartment to model presymptomatic infectiousness. Our basic model is a special case of a so-called $SI^{(n)}R$ multistage infectious model with n denoting the number of infective stages. Typical examples are the well-known SIR model where $n = 1$ and the SEIR model where $n = 2$. Our SEPIR model has $n = 3$ stages and assumes a homogeneously mixed population with five compartments representing those who are (i) susceptible, (ii) exposed but not infectious (iii) presymptomatically infectious, (iv) symptomatically infectious and (v) recovered, as shown in Figure 2.

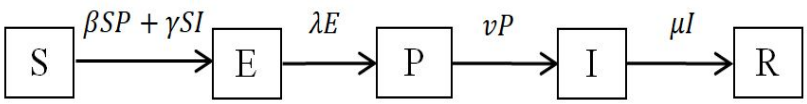

Figure 2. The SEPIR model. The arrow going from S to E represents (i) those infected by someone in P, and also (ii) those infected by someone in I. The three arrows going from E to P, P to I, and I to R represent, respectively, (a) those exposed becoming infectious but not displaying symptoms, (b) those infectious then showing symptoms and (c) those infectious who show symptoms who recover.

In Figure 2 shows the model, with arrows as indicated in the caption The variables *S*, *E*, *P*, *I* and *R* are the number of individuals in each compartment. These satisfy the ordinary differential equations (ODEs):

$$\frac{dS(t)}{dt} = -\beta S(t)P(t) - \gamma S(t)I(t) \quad (1)$$

$$\frac{dE(t)}{dt} = \beta S(t)I(t) + \gamma S(t)I(t) - \lambda E(t) \quad (2)$$

$$\frac{dP(t)}{dt} = \lambda E(t) - \upsilon P(t) \quad (3)$$

$$\frac{dI(t)}{dt} = \upsilon P(t) - \mu I(t) \quad (4)$$

$$\frac{dR(t)}{dt} = \mu I(t) \quad (5)$$

We also include six further parameters t_0, e_0, N, σ, p_R and τ. These are all listed and defined in Table 1; t_0 and e_0 are self-explanatory, and N and σ will be discussed below.

The parameters p_R and τ are not standard. The parameter p_R is the proportion of those who recover well if infected. Because some people can die from COVID-19 the risk of death is of particular concern. The parameter is assumed constant in our model. The population infection fatality rate (IFR) is 1-p_R but an individual's IFR increases rapidly with age. The population IFR can change with time due to: improvements in treatment, changes in immune response due to a new variant, and the amount of stress a health system is subjected to.

We adjust the model to enable the probability p_R to be estimated when fitting the model to data by dividing infectious individuals, all of whom are assumed to 'recover', into two compartments (i) those that recover *well* and (ii) those that *die* due to the virus, as illustrated Figure 3.

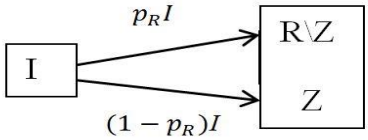

Figure 3. *Adjustment of the SEIR model where R is divided into two compartments, R \ Z, those that recover and Z, those that die; where p_R is the proportion that recover.*

The last parameter τ is a time delay, used to modify equation (5) so that

$$\frac{dR(t)}{dt} = \mu I(t-\tau). \quad (6)$$

Inclusion of τ allows a slightly better model fit to data. Examination of the first wave data from many countries suggest that there is a delay in the hospital reporting of patient recovery. This could be modelled more appropriately by use of a further compartment but as time to recovery is only incidental in our analysis the use of a simple delay as represented by τ is adequate for our purposes.

In the next Section we discuss fitting the model to data. More elaborate models can and have been developed. For example, as already mentioned in the Introduction. Dagpunar (2020), in addition to looking at future waves, extends R into additional compartments in order to represent different outcomes of hospitalization under different interventions.

4 SEPIR MODEL ESTIMATION

4.1 Maximum Likelihood (ML) Estimation of Parameters

We consider first how the SEPIR model can be fitted to observational data. We denote by $\boldsymbol{\theta} = (b1, b2, \ldots, b_m)$ the vector of m parameters. As discussed above we have m = 11 parameters in our SEPIR model. If $\boldsymbol{\theta}$ were already known, the behaviour of the model would be completely specified and the five differential equations (1) to (4) and (6) could then be solved by numerical integration to give the trajectories

$$S(t,\boldsymbol{\theta}), E(t,\boldsymbol{\theta}), I(t,\boldsymbol{\theta}), R(t,\boldsymbol{\theta}), Z(t,\boldsymbol{\theta}) \quad \text{for } t = 1,2,\ldots,M \tag{7}$$

where t is the day and M is the number of days of interest.

Some of the parameters may be known through virological studies, for example, but some may not be. Here we assume that all eleven parameters are unknown and use the standard method of Maximum Likelihood (ML) to estimate all the parameters using observational data obtained from the epidemic. We follow the method outlined in the Supplementary Materials of Dye *et* al. (2020) who applied it to the SEIR model with m = 9 parameters. For clarity, we repeat the description of the method here for the SEPIR model, augmenting the discussion by justifying how the use of bootstrapping can be used to numerically construct confidence intervals for estimates of the trajectories.

Consider first the estimation of all the parameters from a sample of *cumulative* deaths obtained daily over a given period. Let this sample of observed daily cumulative deaths be denoted by

$$\mathbf{Z} = \{z_t \;\; t = 1,2,\ldots M\} \tag{8}$$

where z_t is the total number of deaths up to day t and M is the number of days observed. If the observations were made without error and the parameter values for $\boldsymbol{\theta}$, are correct then the cumulative death trajectory $\{Z(t,\boldsymbol{\theta}) \;\; t = 1,2,\ldots,M\}$ would match the observed cumulative deaths $\mathbf{Z}$ in (8).

The trajectory of an epidemic means that the statistical errors are unlikely to be normally and independently distributed (NID). A better model would be to treat the trajectory as a non-stationary Poisson process so that errors are correlated, or possibly as a non-stationary autoregressive process. We did investigate such models but with a correlation structure not clearly distinguishable from white noise, these did not change our conclusions. The added complication therefore added little and a full discussion would only have obscured the main purpose of the paper. To include statistical uncertainty in the model we therefore assume simply that

$$z_t = z(t,\boldsymbol{\theta}) + e(t) \quad t = 1,2,\ldots,M \tag{9}$$

where $e(t)$ is random error, assumed normally and independently distributed (NID) with standard deviation σ, i.e.

$$e(t) = \; z_t - z(t,\boldsymbol{\theta}) \sim \text{NID}(0,\sigma^2) \tag{10}$$

so that σ is treated as a parameter and is included as a component of $\boldsymbol{\theta}$**.** Statistical variation in the number of persons in each compartment is likely to vary among different compartments. Using different values for some of the compartments had little effect on the parameter values. This is due to the known result that estimates of parameters that are means are statistically mutually independent of the estimate of the variances.

The logarithm of the distribution of the sample is then

$$\text{L}(\mathbf{Z}|\boldsymbol{\theta}) = -\; (M/2)\ln(2\pi) - M\ln\sigma - \; [1/(2\sigma^2)] \sum_{i=1}^{M} [z_t - \; z(t,\boldsymbol{\theta})]^2 \tag{11}$$

where $\mathbf{Z}$ is the random argument and the parameters $\boldsymbol{\theta}$ are fixed. In ML estimation (MLE), this is reversed so that $\mathbf{Z}$ is the known sample of observations now regarded as fixed. We write L as $\mathrm{L}(\mathbf{Z}|\boldsymbol{\theta}) = \mathrm{L}(\boldsymbol{\theta}|\mathbf{Z})$) calling it the (log)likelihood to indicate that it is now treated as a function of $\boldsymbol{\theta}$. The ML estimator $\widehat{\boldsymbol{\theta}}$ is then the value of $\boldsymbol{\theta}$ at which $\mathrm{L}(\boldsymbol{\theta}|\,\mathbf{Z})$ is maximized. i.e.

$$\widehat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\{(\mathrm{L}(\boldsymbol{\theta}|\,\mathbf{Z})\}. \tag{12}$$

Nelder-Mead numerical search (Nelder and Mead, 1965) for the maximum was used. This could be global or local but makes little practical difference given that the fitted epidemic trajectory was visually very good, with the parameter point estimates centrally placed in the bootstrap plots. The search goes through the $\{\boldsymbol{\theta}_i$: i=1, 2, 3, …$\}$, comparing the different $\mathrm{L}(\boldsymbol{\theta}_i, |\mathbf{Z})$ to find $\widehat{\boldsymbol{\theta}}$, the best $\boldsymbol{\theta}$.

The method may be extended not only to fitting to deaths data but to include other data samples such as

$$\mathbf{Y} = \{y_t \;:\; t = 1,2, \dots, M\} \tag{13}$$

where $\boldsymbol{y_t}$ is the number of prevalent active cases on day t. Fitting simultaneously to both $\mathbf{Y}$ and $\mathbf{Z}$ can be carried out by adding to the right-hand side of (11) a corresponding set of terms for $\mathbf{Y}$

Each step of the Nelder-Mead optimization is as summarized by Cheng (2017, Equation 3.27) and is as follows. We start the Nelder-Mead search by choosing initial values for all the parameters:

$$\boldsymbol{\theta}^{(0)} = [\beta^{(0)}, \gamma^{(0)}, \lambda^{(0)}, v^{(0)}, \mu^{(0)}, (t_0)^{(0)},\ (e_0\,)^{(0)}, N^{(0)},\ \sigma^{(0)}, (p_R)^{(0)}, \tau^{(0)}].$$

This builds a simplex of (m + 1) parameter points $\boldsymbol{\theta}_i^{(1)},\ \mathrm{i} = 1, 2, \dots, m+1,$ one of which is typically $\boldsymbol{\theta}^{(0)}$. We then calculate the loglikelihood $\mathrm{L}(\boldsymbol{\theta}_i^{(1)}|\mathbf{Z})$ at each parameter point of the simplex and compare their values, adjusting the position of one or more of the simplex points to improve the loglikelihood. The process is repeated until the simplex converges on a maximum; typically shrinking the dimensions of the simplex until all the points are clustered round the maximum point. Details of the precise methodology are given in (Nelder and Mead, 1965).

To calculate the loglikelihood at a particular point, $\boldsymbol{\theta}$, we first calculate the trajectory of each of the variables, *S*, *E*, *P*, *I*, *R*, as given in Equation (7). We used Euler step-wise integration of the differential equations (1) – (4) and (6) with a step-length of 1/8th of a day. The differential equations are scale invariant so we can standardize the equations by taking

$$(S + E + P + I + R) = 1.$$

This converts all five variables into fractions, and is done for convenience in calculating the trajectories.

The initial values of the trajectory at t = 0, are $S(0,\boldsymbol{\theta}\,) = 1,\ P(0,\boldsymbol{\theta}\,) = I(0,\boldsymbol{\theta}) = R(0,\boldsymbol{\theta}\,) = 0)$ for all the simplex points. Calculating the trajectories involves $\beta, \gamma, \lambda, v, \mu\ \ t_0,\ e_0$ and τ but not $N,\ \sigma$ or p_R. However, the loglikelihood expression given in (11) does involve σ. Moreover, $z(t, \boldsymbol{\theta})$ needs to be the cumulative number of *actual* deaths as given by the Z trajectory, and not a standardized version of *Z*. This can be calculated as

$$z(t, \boldsymbol{\theta}) = (1 - p_{\mathrm{R}})NR(t, \boldsymbol{\theta}). \tag{14}$$

Thus, all the parameters will be adjusted by Nelder-Mead in calculating the loglikelihood.

4.2 Switzerland: A Numerical Example

Table 1 shows the result of fitting SEPIR using M = 109 days of data based on daily observations in Switzerland starting on 15 Feb 2020 for two series: Daily New Cases and Daily Deaths. All 11

parameters were estimated by ML. The values, in particular, of two parameters inform concerns about the effect of the epidemic on the population.

Firstly, because the differential equations are scale invariant, the population size, *N*, can be standardised to unity. But we treat *N* as the unknown size of the population actually at any risk of exposure to the virus, allowing for the possibility this might not include the whole population. The estimated population size of $\widehat{N} = 35{,}000$ is small compared to the actual population size of 8.2 million. The main reason for the difference is that the model does not include a mechanism of epidemic control which we know took place in every country to prevent the spread of infection. The SEPIR model, which assumes a homogeneously mixed population can only allow for this by changing the population size, indicating that lockdown or self-isolation drastically reduced the number of persons at any risk of exposure to the virus. Without examining regional records, a further contributing reason for the estimated small size of the population exposed may be that the outbreak in Switzerland was mainly confined to parts nearest Italy, the first European country to be badly affected by COVID.

Table 1. *Parameters of the SEPIR model with estimates for Switzerland*.

Symbol	Definition	Estimated Value	95% CI
β	presymptomatic transmission rate	0.45	(0.37 – 0.53)
γ	symptomatic transmission rate	0.15	(0.04 – 0.240)
λ^{-1}	mean latent period in compartment E (days)	1.2	(0.34 – 1.43)
v^{-1}	mean presymptomatic period (days)	4.3	(2.23 – 5.70)
μ^{-1}	mean symptomatic period (days)	14.6	(11.49 – 15.32)
t_0	number of days from start of epidemic before observations began	30.5 days	(23.1 – 36.3)
e_0	initial proportion of individuals exposed	4.0 E-07	(1.2E-7 – 2.0E-6)
N	numerical size of exposed population	3.50E+4	(3.37 – 4.80)E+4
σ	standard deviation of observational error	564	(484 – 646)
p_R	probability of an infective recovered well	0.942	(0.936 – 0.956)
τ	mean duration spent in R state	2.8 days	(0.17 – 3.81)

The other parameter that we can immediately highlight is p_R, the probability of recovering well if one is infected. We have the estimate $\hat{p}_R = 0.942$ which is lower than that which is now currently accepted as being the case for a population as a whole. However, it is well known that the probability of dying from COVID-19 increases roughly exponentially with age and those infected in Switzerland at the start of the epidemic were older so this might be why the estimate of p_R is low.

The usefulness of the SEPIR model in investigating presymptomatic transmission is demonstrated in that the mean number of infections caused by the infector is simply the product of the presymptomatic transmission rate, β and the mean presymptomatic period, v^{-1}, that is βv^{-1}. Likewise the mean number of symptomatic infections caused by the infector is $\gamma\mu^{-1}$. The proportion of presymptomatic infections is therefore $(\beta v^{-1})/(\beta v^{-1} + \gamma\mu^{-1}) = P_{pi}$, say. Using the estimates of the parameters from Table 1, we obtain the ML estimate of this proportion using the ML estimates of the parameters

$$\hat{P}_{pi} = \left(\hat{\beta}\hat{v}^{-1}\right)/\left(\hat{\beta}\hat{v}^{-1} + \hat{\gamma}\hat{\mu}^{-1}\right) = \frac{0.45\times4.3}{(0.45\times4.3+0.15\times14.6)} = 46.9\%. \quad (15)$$

This value is similar to the value of 44% found by He et al. (2020).

5 FITTED TRAJECTORIES AND CONFIDENCE INTERVALS

5.1 Swiss First Wave Trajectories

Once ML estimates of the parameters have been obtained, fitted trajectories can be obtained by substituting these estimates into Equation (7). Figure 4 charts the Daily Active Cases and the Cumulative Deaths trajectories for the Swiss data together with their corresponding data. It will be seen that visually the fits are good.

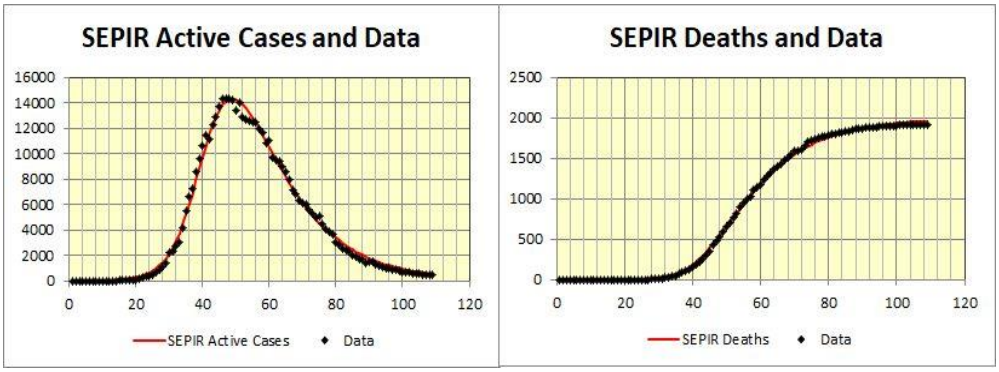

Figure 4: SEPIR Active Cases and Cumulative Deaths fitted to Swiss Data. Horizontal axis is days with day #1 = 15th February 2020.

Visual comparison of the fitted trajectories and the observed data is quick but informal. To be useful, one would want to be sure that parameter values can be found so that SEPIR models the behaviour of the epidemic sufficiently accurately for practical purposes. We adopt the usual criterion which is that the parameter estimates be statistically consistent. Under mild conditions, parameter estimates obtained by ML estimation are consistent, see for example Cox and Hinkley (1974, Section 9.2 ii). Our first step is therefore to check numerically that the ML parameter estimator $\widehat{\boldsymbol{\theta}}$ gives a local maximum of the loglikelihood. With normal errors, as we have assumed in Equation (9), a simple practical way of seeing that we have a local maximum, is to numerically check that the negative of the Hessian matrix of second derivatives of the likelihood is positive definite (Beck 2018, Theorem 2.27b). This can be done by numerically checking that the eigenvalues are all positive. Bootstrapping works well in this context in providing a way to calculate not only confidence intervals for the parameter estimates but for the trajectories a well.

Confidence intervals for the estimated parameter values can be obtained using the bootstrap procedure described in Subsection 4.1.3 of Cheng (2017) and, for confidence intervals for points on a trajectory (this being a function depending on $\boldsymbol{\theta}$), using the procedure in Subsection 4.2 of Cheng (2017). In summary the bootstrap procedure is as follows. If the true parameters were known, we could in principle numerically produce B sets of data 'observations' $\{\mathbf{Y}_k, \mathbf{Z}_k\}$ $k = 1, 2, \ldots, B$, where k is the kth set of observations with Y_k as in equation (13) and Z_k as in Equation (8) and errors satisfying Equation (9) so that each set of observations has the assumed model form. If we then fit the model to each sample using ML, each resulting fit would look like Figure 4, but have the same inherent statistical variability arising from the same form of statistical error as in Equation (9) with the errors between sets mutually independent. Estimates of statistical error can therefore be obtained from the sample $\{\mathbf{Y}_k, \mathbf{Z}_k\}$ $k = 1, 2, \ldots, B$.

We do not have the true parameter values of course. But the bootstrap principle says that we can replace them by the best values to hand, which are the ML estimates of the parameters obtained from the original sample. The now vast literature confirms that this apparent 'something from nothing' principle works well. In fact, it has a *better* asymptotic performance than classical asymptotic theory in that the asymptotic probability distributions of the estimates of the parameters obtained by bootstrapping converge to the same distributions as those given by classical asymptotic theory but

they converge *faster*. Here the term 'asymptotic' refers to the size of the original data sample, in the case of our problem this is *M* and has nothing to do with *B*, the number of bootstraps, which affects the accuracy of estimating σ, the statistical error of the ML estimates. This accuracy can be made as precise as we like by making *B* sufficiently large and is unconnected with σ.

As explained in Subsection 4.1.3 of Cheng (2017), calculation of the confidence interval (CI) of a given parameter estimate is obtained simply by ranking the values of the given parameter in the bootstrap sample and taking appropriate percentiles of the ranked values as the upper and lower limits of the CI. This is the way the CIs displayed in Column 4 of Table 1 have been obtained for the estimated parameters of the Swiss data. A visual bonus of calculating confidence intervals using bootstrapping is that we can examine the joint distributions of *pairs* of parameter estimates using their bootstrap scatterplots. These scatterplots would correspond to the bivariate normally distributed points if *M* is sufficiently large for these distributions to be near their asymptotic limits.

The upper right set of ten plots in Figure 5 shows the scatterplots corresponding to all the possible combinations 500 bootstrap estimates of pairs of the virus transmission and state duration parameters: $\beta, \gamma, \lambda^{-1}, v^{-1}, \mu^{-1}$ (labelled as b1 to b5 in the Figure) obtained for the Swiss data. These plots show that the estimates are fairly normally distributed and have sufficient axial symmetry to indicate no strong correlation between any pair. The only concern is that the positions of some of the ML estimates are asymmetrically placed relative to the bootstrap scatter. This may mean that the sample size of the data is insufficiently large for asymptotically normality theory to be fully approached.

Correlation between the parameter estimates can result in overfitting, this affecting estimation of the pdf of the serial interval. However, our main interest is in the proportion of presymptomatic infections P_{pi} which depends only on β, γ, v^{-1} and μ^{-1}. The scatterplots of Fig. 5 indicate that there is little correlation between these parameters, so the estimate of P_{pi} should not be overfitted.

The lower left set of three scatterplots in Figure 5 corresponds to the pairs involving the parameters μ^{-1}, N, p_R (labelled b5, b8, b10 in the Figure 5). These parameters *are* strongly correlated, with bootstrap sample correlation coefficients $r_{5,8} = -0.94$, $r_{5,10} = -0.81$, $r_{8,10} = 0.80$. The red dots in the scatterplots correspond to the ML estimates of parameters. Their positions show that confidence intervals with equal tail probability regions are not symmetrically centered relative to these ML values.s None of the scatterplots of other pairs show noticeable correlation and are not shown.

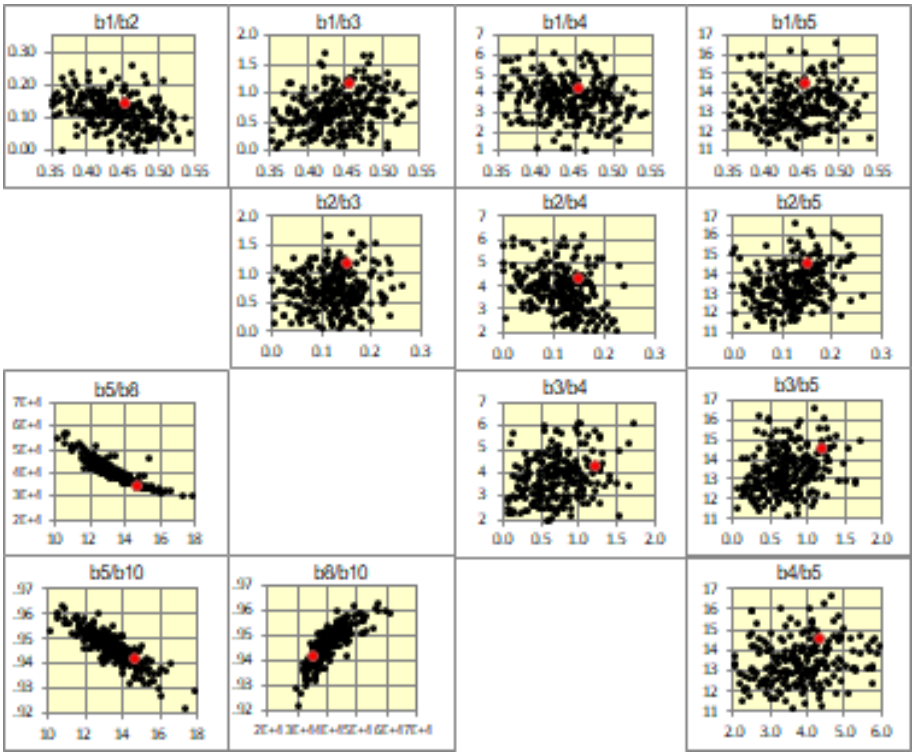

Figure 5: Selected scatter plots of 250 pairs of bootstrap parameter estimates for Swiss data. (500 bootstraps were carried out but for clarity only half are displayed. Upper right set of ten plots show all pairs of the parameters $\beta, \gamma, \lambda^{-1}, \nu^{-1}, \mu^{-1}$ (labelled b1, b2, b3, b4, b5 in the Figure). Lower left set of three plots show all pairs of the parameters μ^{-1}, N, p_R (labelled b5, b8, b10 in the Figure). Red dots are the ML estimates of the parameters

It will be seen that there is a strong correlation between μ^{-1}, N, p_R indicating that the negative of the Hessian matrix for second derivatives is not strongly positive definite so that we do not have a clear local maximum of the likelihood. Geometrically, high correlation between two variables means that one or other variable can be changed provided the other is altered to compensate, so that the (optimized) value of the likelihood remains unchanged to first order. With three variables, a three-way coordinated adjustment is possible without significantly changing the likelihood value.

We have not systematically assessed the kind of adjustments that can be made without changing the value of the maximized value as our main interest is in assessing the proportion of presymptomatic infections. We therefore simply note the confidence limits for the ML estimate given in Equation (~~14~~ 15) of the probability, $\hat{P}_{pi} = 46.9\%$, of someone being infected presymptomatically. The lower and upper limits can be calculated as the corresponding percentiles of the bootstrap sample of P_{pi} estimates giving the CL as (26.8% , 78.7%). Scatterplots of $\hat{P}_{pi}$ versus each of the four parameters β, γ, ν^{-1} and μ^{-1} on which it depends are shown in the Appendix.

The confidence level applies to the trajectory at a single particular point of the trajectory. This method has been used for the trajectories in Figure 6 where, for simplicity of presentation, the upper and lower limits of the confidence interval are depicted as bands, but the values should only be applied to a single day. If applied to more than one point simultaneously the level of confidence is lowered by an unspecified amount.

However, it is possible to calculate the green and blue curves so that the confidence level can be applied to the whole curve *simultaneously*, so that the whole of the MLE trajectory (shown in red) lies entirely between the green and blue curves with given level of confidence. The underlying theory and methodology are given in Subsection 4.3 of Cheng (2017) where the method described is actually conservative with the confidence level no less than the stipulated level.

Figure 6 shows the fitted SEPIR cumulative deaths trajectory (red) together with the observations (black). The trajectories in grey are the bootstrap cumulative deaths trajectories, each calculated from a bootstrap sample in the exactly same way as the calculation of the trajectory fitted to the original sample. Only 250 bootstrap trajectories are shown because of charting limitations, however the 95% confidence limits shown in green and blue are calculated from all 500 bootstrap samples.

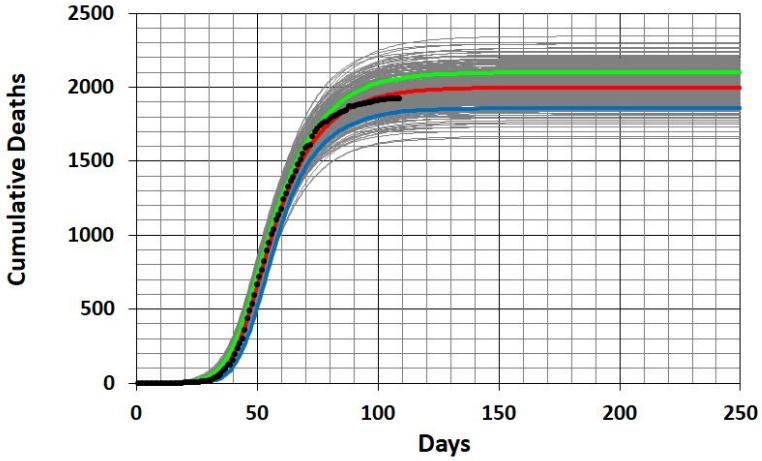

Figure 6: *SEPIR Fitted Cumulative deaths trajectory (red) for Swiss data obtained from 109 observations (black). Upper (green) and lower (blue) confidence limits.*

6 R_T THE EFFECTIVE REPRODUCTION NUMBER

The SEPIR model, like the well-known SIR and SEIR, is an SI$^{(n)}$R model, where there are *n* multiple infectious stages (with *n* = 3 in the case of SEPIR), as defined in Equations 6a to 6d in Ma and Earn (2006). Their Equation (7) gives formulas for the Reproduction Number, R_0, in SI$^{(n)}$R models. (The Reproduction Number R_0 is simply, but precisely, defined as the number of susceptible individuals that an infectious person will go onto infect when the epidemic first starts, assuming that the population is homogeneously mixed.) For the SEPIR model, using the parameters that already appear in Equations (1)-(4), we have

$$R_0 = \beta/\nu + \gamma/\mu\ . \tag{16}$$

In the case of Switzerland, using the estimated values of the parameters in Table 1, we have

$$R_0(Switzerland) = 0.45 * 4.3 + 0.15 * 14.6 = 4.13 \tag{17}$$

which seems reasonable.

As mentioned in the Introduction, in practice R_t, the effective reproduction number, is more useful as it can be continually used to gauge how well control strategies are working throughout the epidemic. The theoretical basis underlying the calculation R_t is described by Ma (2020). We have

$$R_t = \frac{c(t)}{\int_0^\infty c(t-u)w(u)du}, \tag{18}$$

where $c(t)$ is the incidence curve of new cases at time t and $w(u)$ is the serial interval probability density function (pdf) of a secondary case. The denominator in Equation (18) measures how the new cases at time t arise from those infected prior to time t. The epidemic is rising or falling depending on whether the numerator is larger or smaller than the denominator, with equilibrium when they are equal. Thus R_t has the critical reproduction property of R_0 but moreover is dynamic, so that it can be used to gauge the progress of the epidemic as this develops.

It turns out that the formula (18) is robust so that the serial interval distribution does not have to be estimated all that accurately. In fact, Germany, early during its first COVID-19 epidemic wave, used the simple denominator $c(t\text{-}4)$. Cori *et al*. (2013), using a Bayesian approach, examined various empirically obtained serial interval distributions drawn from different epidemics. Dye *et al*. (2020) used a discretized and shifted gamma distribution $g(t),\ t = 1,2,\ldots,12$ to represent the serial interval distribution $w(u)$ that is shown as the dashed curve in Figure 1, calculating the denominator as

$$D = \sum_{u=1}^{12} c(t-u)g(u). \tag{19}$$

Figure 7 depicts R_t calculated using this formula for Switzerland when c(t) is a daily 7-day moving average of new cases.

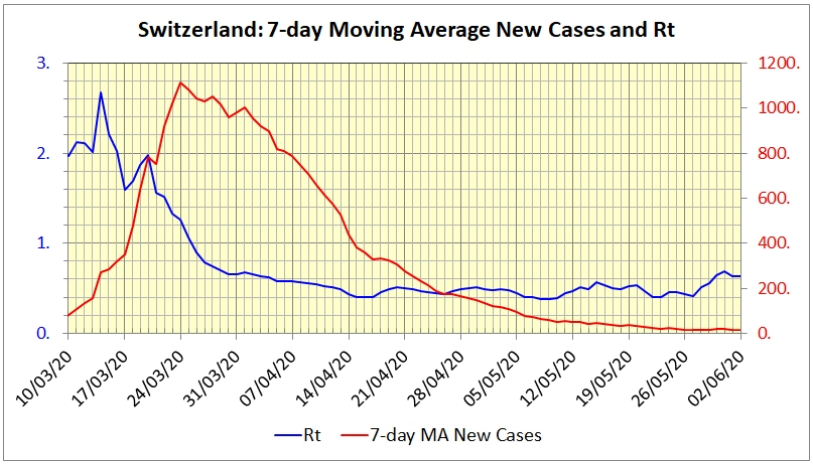

Figure 7: Chart of the effective R_t calculated using the formula in Equation (19) for Switzerland where $c(t)$ is the 7-day moving average of new cases.

An important point is that how the epidemic ultimately ends depends on R_0, not R_t. Until the arrival of vaccines all controls, lockdown, hand-washing, social distancing and so on, affect only β, moreover only temporarily. Thus R_t, which depends on this temporary β, varies as the controls vary, and so is a simple gauge of how current controls are doing. However once controls are removed, β returns to its original value as in R_0, so that the epidemic returns, causing another wave. As determined by Kermack and McKendrick (1927) for the SIR model, the ultimate end, when a given number of susceptibles have been infected, is determined by R_0; the remaining uninfected susceptibles, $S(\infty)$, satisfying the so-called final-size relation:

$$\ln [S(0)/S(\infty)] = R_0[1 - S(\infty)/S(0)]. \qquad (20)$$

This shows how $S(\infty)$/S(0) depends purely on R_0. The equation is implicit but is easily solved numerically. Figure 8 charts $S = S(\infty)/S(0)$ with $S(0)$ set equal to unity, which we can do so without loss of generality as Equation (20) shows that S is scale invariant.

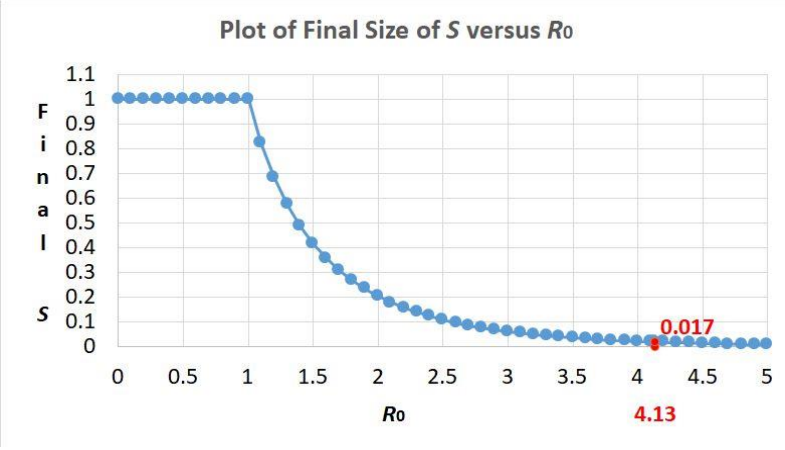

Figure 8: Plot of Final size of S versus R_0

We have remarked above that Ma and Earn (2006) give an explicit formula for calculating R_0 in SI$^{(n)}$R multistage infectious models which applied to our estimated Swiss data example gives $R_0 = 4.13$ as in Equation (17). They show that the final size Equation (20) also applies to SI$^{(n)}$R models, and so to the SEPIR model. This gives $S(\infty) = 0.0172$, and this is highlighted in Figure 8, showing that over

98% for the population would need to have been infected before the Swiss epidemic ends. Relying on the final size relation to end the epidemic is not a strategy any country can count on.

Vaccination (assuming 100% efficacy and no waning of immunity) works differently by moving susceptibles directly to the R compartment so that the effective population shrinks. This is most clearly seen if immunization takes place before the epidemic so that a proportion p of the population is immunized (Brauer *et al*. 2019). Then R_0 would fall to $R_0(p) = \beta_B N(1-p)/\mu$. Here β_B is the β in Brauer *et al*. 2019 which assumes mass action incidence where an infectee makes $\beta_B N$ contacts sufficient to transmit infection in unit time.) $R_0(p)$ is less than 1 if

$$p > p_c, \text{where } p_c = 1 - 1/R_0,$$

Figure 9 charts the graph of p_c versus R_0 when there would be no epidemic. In the case of Switzerland, where in Equation (17) we estimated from the first wave that $R_0 = 4.13$, Figure 9 shows that, assuming that a vaccine gives an individual 100% protection against infection and eliminates transmission given exposure, then over three-quarters of the population need vaccination to stop the epidemic. Thus, a sizeable but not all of the population needs immunization to prevent the epidemic. In fact, current vaccines are of the order of 70-90% efficacious and preliminary data suggests they may be able to reduce transmission by perhaps 50%

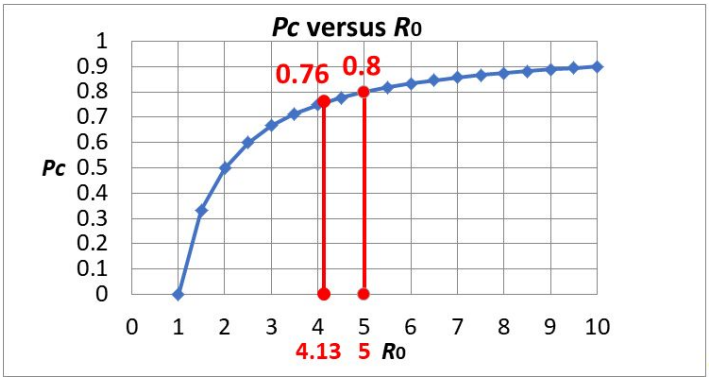

Figure 9: Chart of p_c versus R_0

The only long-lasting control is vaccination, At the time of writing, the rate of vaccination in the UK seems sufficiently effective to steadily reduce the epidemic whilst allowing relaxation of the restrictions on social mixing. However, there is concern over the long-term effectiveness of existing vaccines against the global rise of mutant strains in the virus that are inevitably occurring.

New variants likely to be imported suggest that a basic reproduction number of R_0 as high as 5 is not out of the question. This would mean, as shown in Figure 9, that without Non-Pharmaceutical Interventions, 80% of the population would need vaccination before the growth rate became negative. Bearing in mind that vaccination is not 100% effective and that a significant proportion of the population do not wish to be vaccinated anyway, it seems likely that Covid-19 will become endemic with regular annual booster vaccination and test, trace, and isolation needed in the long term.

7 CLOSING REMARKS

7.1 Caveats

Firstly, we have not examined in detail the robustness of the maximum likelihood optimization used to fit the model. In our numerical example we chose the first wave of the epidemic in Switzerland because the data corresponded well to the characteristics of the SEPIR model. However even in this example, alternative good fits can be achieved with combinations of parameter values different from those reported in Table 1. Thus, in practice, where possible, comparison with parameter estimates obtained in other ways should be made, in order to assess when the estimates obtained can be relied

upon ~~on~~. Our study focuses on the first wave, but an epidemic like Covid-19, that is not immediately controlled, recurs with many further waves. Modelling of such multi-wave behaviour is therefore of interest, but is not discussed here.

Secondly, the simplicity of models such as SEIR or SEPIR means that the practical usefulness of using them on their own is limited. The models are idealizations of the way the epidemic behaves and of population behaviour. Thus, control policies are not modelled nor their influence on population behaviour. Indeed lack of homogeneity of population behaviour is an important factor that has to be addressed in implementing control policies because these latter have to recognize the issues they give rise to for the population as a whole to be willing to follow them At the time of writing, there has been a resurgence of the virus and more virulent virus strains have appeared but their impact has been balanced by the availability of vaccines This all requires a national control policy which is fair. On this basis the present national policy to go in "earlier and harder and stay longer than might be thought necessary" even in areas with low prevalence it seems~~.~~ Ideally a detailed model allowing for local differences is required but this seems unrealistic given the speed of changing events. However, less complicated models like the present SEPIR model may be helpful in informing decision.

It should be noted that our transmission rates, though estimated, are assumed constant rather than time dependent which would be needed to model changing management of the epidemic. This could explain why the estimates of some of the biological parameters are rather different from those found in some other studies. We reiterate that our model does not include all aspects of the behaviour of Covid-19. Nevertheless, it is interesting that, though our data has a simple form, it allows presymptomatic behaviour to be quantified.

We repeat one final aspect of Covid-19, mentioned in the Introduction, that is of interest, namely, that 'presymptomatic' is not the same as 'asymptomatic'. This could be handled by including an extra compartment to represent those people who are asymptomatic. However, detailed hospital data would be needed to allow the proportions in each compartment to be estimated, together with a more involved analysis than we have used in this paper. We have therefore not included such aspects as their inclusion would possibly detract the reader's attention from the main purpose of our paper. Presymptomatic behaviour seems sufficiently interesting and important to be considered in its own right.

7.2 Summary

The SEPIR model is an extension of the well-known SEIR model. Both are particular cases of the more general $SI^{(n)}R$ model with multiple infectious stages. For Covid-19, the SEPIR compartment, P, is used to represent those infected presymptomatically. Compared with the SEIR model this is an important improvement, as shown here. This latter is based on recorded data, and clearly shows the large part played by presymptomatic infection in the case of Covid-19. The practical implication is that Covid-19 control strategies need to recognise and deal with presymptomatic transmission to be truly successful. Our formulation of the SEPIR model includes adjustable parameters which can either be given or fitted to observed data. An important aspect of our parametrization is that it allows estimation of an initial susceptible proportion that is less than unity rather than assuming that the susceptible people comprise the country's whole population, as is usually assumed. This relaxes the assumption of homogeneity of virus transmission throughout the whole population, as this assumption may not be reasonable, especially in the early stages of an epidemic where the number of individuals infected is initially small.

REFERENCES

Beck A. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, Philadelphia. https://archive.siam.org/books/mo19/MO19_ch2.pdf

Brauer F, Castillo-Chavez C, Feng Z. (2019) Simple Compartmental Models for Disease Transmission. In: *Mathematical Models in Epidemiology. Texts in Applied Mathematics*, vol **69**. Springer, New York, NY. https://doi.org/10.1007/978-1-4939-9828-9_2

Cheng R C H (2017). *Non-Standard Parametric Statistical Inference*. Oxford University Press, Oxford.

Cori A, Ferguson N M, Fraser C and Cauchernez S (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology* **178(9)**: 1505-1512.

Cox D R and Hinkley D V (1974). *Theoretical Statistics.* Chapman and Hall, London.

Dagpunar J S (2020). Sensitivity of UK Covid-19 deaths to the timing of suppression measures and their relaxation. *Infectious Disease Modelling* **5:** 525-535 https://doi.org/10.1016/j.idm.2020.07.002

Dye C, Cheng R C H, Dagpunar J and Williams B G (2020). The scale and dynamics of COVID-19 epidemics across Europe. *R. Soc. Open Sci. .* **7**: 201726 https://doi.org/10.1098/rsos.201726

He L, Lau E H Y, […], Leung G M (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine* **26**: 672-675.

Kermack W O and McKendrick A G (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A* **115**: 700-721.

Ma J and Earn D J D (2006). Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bulletin of Mathematical Biology* **68**: 679-702.

Ma J (2020). Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling* **5**: 129-141.

Nelder J A and Mead R (1965). A simplex method for function minimization. *Computer Journal*, **7**: 308-313.

Tapiwa G, Kremer C, Chen D, Torner A, Faes C, Wallinga J and Hens N (2020). Estimating the gerneration interval for coronavirus disease (COVID-19) based on symptom onset, March 2020. *Euro Surveill*. **25(17)**: pii=2000257.

APPENDIX

Figure A below shows the scatterplots of $\hat{P}_{pi}$ versus each of the four parameters β, γ, v^{-1} and μ^{-1} on which it depends.

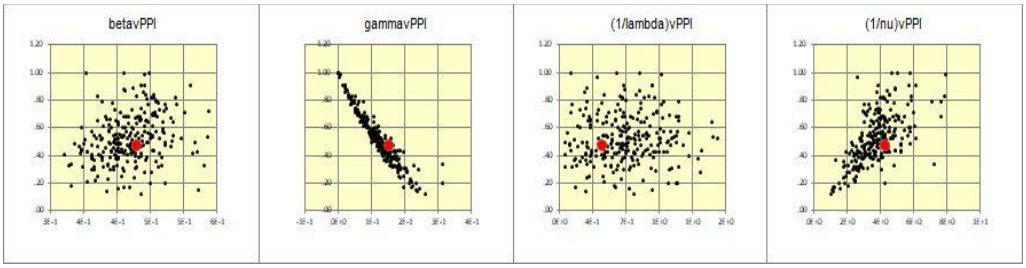

Figure A: Selected scatter plots of the bootstrap estimates of P_{pi} versus the four parameters β, γ, v^{-1} and μ^{-1}on which it depends. The red dots are the ML estimates of the parameters

AUTHOR BIOGRAPHIES

RUSSELL CHENG retired from the University of Southampton in 2007 where he had been Head of the Operational Research Group, having held previous positions at Cardiff University and the University of Kent at Canterbury. https://www.southampton.ac.uk/maths/about/staff/rchc.page

627
628 **CHRISTOPHER DYE** FRS, FMedSci, has held positions at the London School of Hygiene and
629 Tropical Medicine, the World Health Organization and Gresham College London. He is currently
630 Professor of Epidemiology at Oxford University http://en.wikipedia.org/wiki/Christopher_Dye
631
632 **JOHN DAGPUNAR** retired from Edinburgh University in 2008. He is Visiting Research Fellow in
633 Mathematical Sciences at the University of Southampton. His research interests are in simulation,
634 financial mathematics, health studies, and reliability.
635 https://www.southampton.ac.uk/maths/about/staff/jd2y15.page
636
637 **BRIAN WILLIAMS** is Senior Research Fellow at the South African Centre for Epidemiological
638 Modelling and Analysis (SACEMA) having held the position of Epidemiologist at the World Health
639 Organization from which he retired in 2008.