Biometrics 000, 000–000 DOI: 000
000 0000

Web-based Supplementary Materials for A Path-Specific SEIR Model for use with General Latent and Infectious Distributions by Aaron T. Porter and Jacob J. Oleson

Aaron T. Porter*

Department of Biostatistics, University of Iowa, 52242, USA

**email:* aaron-t-porter@uiowa.edu

and

Jacob J. Oleson*

Department of Biostatistics, University of Iowa, 52242, USA

**email:* jacob-oleson@uiowa.edu

1. Web Appendix A

The PS SEIR model can be derived as a stochastic analog to the following nonlinear system of ordinary differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= -f(\underline{\psi}, t)S\frac{I}{N};\\
\frac{dE}{dt} &= f(\underline{\psi}, t)S\frac{I}{N} - g(\underline{\alpha}, E);\\
\frac{dI}{dt} &= g(\underline{\alpha}, E) - h(\underline{\gamma}, I);\\
\frac{dR}{dt} &= h(\underline{\gamma}, I).
\end{aligned}$$

Several assumptions are made in this process, and we outline the core assumptions here.

1. Assume a homogeneous population with regards to susceptibility. This is commonly assumed in population averaged models.

2. Assume independent Poisson contact distributions for infectious individuals, all of which share a single parameter. This works well for diseases such as mumps or measles, but works poorly in models for sexually transmitted diseases, such as gonorrhea or chlamydia.

3. Define the Exposed compartment as only containing those who will eventually become infectious, and do not consider the possibility of a return to the Susceptible class.

4. Assume constant infectivity throughout the course of the infectious process.

5. Assume independent probabilities of moving from the Exposed Compartment to the Infectious Compartment (as well as from the Infectious Compartment to the Removed Compartment).

6. Individuals are treated as having identical latent and infectious time distributions. There is no individual heterogeneity in these processes.

The full derivation can be found in the Web Appendix. Suppose $g(\underline{\alpha}, E) \geqslant 0$ and the equation $\frac{dE}{dt} = -g(\underline{\alpha}, E)$ has the solution $E = G(\underline{\alpha}, C, t)$ where C is the constant of integration. If there exists a C^* such that $\frac{E}{E_0} = F(\underline{\alpha}, C^*, t)$ where E_0 is the total number of individuals who will become exposed throughout the epidemic, and the conditions $F(\underline{\alpha}, C^*, 0) = 1$ and

$F(\underline{\alpha}, C^*, \infty) = 0$ are met, then $F(\underline{\alpha}, C^*, t)$ is a survival function, and a path-specific analog can be found to the deterministic system under consideration.

Given that an individual is in the Exposed compartment at time t, the probability of a compartmental shift to I before time $t + h$ is $\frac{F(\underline{\alpha},C^*,t)-F(\underline{\alpha},C^*,t+h)}{F(\underline{\alpha},C^*,t)}$. To discretize the above nonlinear system, partition the time over which the epidemic occurs into regular blocks of length h. Then $F(\underline{\alpha}, C^*, t)$ can be approximated by $\prod_{k=0}^{j} \frac{F(\underline{\alpha},C^*,kh)-F(\underline{\alpha},C^*,(k+1)h)}{F(\underline{\alpha},C^*,kh)}$ for t in the interval $((jh, (j + 1)h)$. This places a point mass for the discretization on the left endpoint of the interval, which is consistent with the discretization of the binomial process leading to new exposures below.

Note that $\frac{F(\underline{\alpha},C^*,j)-F(\underline{\alpha},C^*,j+h)}{F(\underline{\alpha},C^*,j)} = P(Z_1 \leqslant j + h|Z_1 > j)$. Consider breaking the Exposed compartment into M_1 distinct bins. In bin j, there is a constant probability, $P(Z_1 \leqslant j + h|Z_1 > j)$, of a compartment change. Using i.i.d. Bernoulli random variables to accommodate the process of a random compartmental change for each individual in bin i, we see that the number of individuals moving from $E_{i,j}$ into $I_{i+1,1}$ is distributed as $binomial(E_{ij}, I_{i+1,1}) = binomial(E_{ij}, P(Z_1 \leqslant j + h|Z_1 > j))$. Assuming that the number of individuals experiencing compartmental change is independent of the bin after conditioning, we retrieve the path-specific process for the latent times found in Equation 2 of the paper. Note that the same process will yield the path-specific process for infectious times.

Next, we derive the binomial process leading to new exposures. The derivation is very similar to the one found in Mode and Sleeman (2000), with the difference being that we use discretized Poisson contact rates and probabilities of contracting infections from infectious individuals. This does not change the overall flow of the derivation found in their book, but our additions are informative to the properties of our model, and so we include the derivation here.

Assume contacts are made at a rate $P(C(t) = c) = x(c, t)$, where $C(t)$ is a random variable indicating the number of contacts an individual makes at time t, and c is a realization of this random variable. Assuming independent contacts, the probability of escaping infection between time t and time $t + h$ is $Q(t) = \sum_{c=0}^{\infty} x(c, t)q^c(t)$, where $q(t)$ is the discretized probability of not developing an infection based on a contact with a single individual in this interval. Assume contacts are made according to a process that can be approximated discretely by a set of T Poisson random variables with rates $\lambda(t)$, which may vary over the T time points. This will allow for differences in contact rates due to interventions. Additionally, denote the probability of contracting a latent infection based on contacting a random member of the population as $q(t)$, where $q(t)$ has been discretized into T values. Then, for a fixed time point t, we have $Q(t) = \sum_{c=0}^{\infty} \frac{\exp(-\lambda(t)h)(\lambda(t)h)^c}{c!} q^c(t)$, implying $Q(t) = \exp(h\lambda(t)(q(t) - 1)) = \exp(-h\lambda(t)p(t))$.

In order to include public health style interventions into our model, assume $p(t) = p^*(t)I(t)/N$, where $p^*(t)$ is the time dependent probability of contracting an infection based on contacting a single infectious individual from time t to time $t+h$, and is discretized into T values. Note that we have parameterized $\lambda(t)p^*(t)$ as $f(\underline{\psi}, t)$ because there is only data to estimate one parameter in the case where there are no interventions. Therefore a single mixing parameter is typically used in these cases, rather than the product $\lambda(t)p(t)$.

Finally, approximate the mixing process $f(\underline{\psi}, t)SI$ by discretizing it and placing a point-mass at the left endpoint of the interval $(ih, (i+1)h)$. This yields the probability distribution of the number of new latent infections at time $i + 1$, given the state of the system at time i, as $binomial(S_i, 1 - \exp(-f(\underline{\psi}, i)h\frac{I_{i+}}{N}))$. This completes the derivation.

As an example of this process of determining the conditional probabilities for the path-specific process through the exposure matrix, consider the standard case of exponentially distributed latent times. Anderson and May (1991) shows that the form of $\frac{dE}{dt}$ is:

$$\frac{dE}{dt} = f(\underline{\psi}, t) - \alpha E.$$

In deriving the survival function, one only needs to consider $\frac{dE}{dt} = -\alpha E$. Solving this equation yields $E(t) = C * \exp(-\alpha t)$, where C is determined by the initial conditions of the system. In order to determine the survival function, we consider $E(t = 0) = E_0$, where E_0 is the total count of individuals who will move into the Exposed compartment over the course of the epidemic. The survival function is then $\frac{E}{E_0} = \exp(-\alpha t)$, which is can be identified as the survival function of an exponential random variable. If we call this random variable Z_1, the discretization we consider is $P(Z_1 < t + h | Z_1 > t)$. This yields $\exp(-\alpha h)$, which is constant in t, and is of the form commonly seen in the SEIR model literature.

2. Web Appendix B

In this section we propose an efficient sampling scheme for the exposure matrix, and discuss the properties and limitations of the sampling scheme.

2.1 *The Sampling Scheme*

We recommend the following update scheme for the exposure matrix: 1) Select a time at which an individual moved to the Infectious category. 2) Select, at random, a path that corresponds to this infection time. 3) Remove this path. 4) Select a new starting time for the exposure path, with equal probability placed on every day between one day before the transfer to the infectious compartment, and M_1 days before the transfer. 5) Add this path to the exposure matrix, and keep this update if $\frac{\Pi(\mathbf{E'})}{\Pi(\mathbf{E})}$ is greater than a randomly generated uniform random variable where $\Pi(\mathbf{E'})$ is the likelihood corresponding to the new exposure matrix, and $\Pi(\mathbf{E})$ corresponds to the likelihood of previous exposure matrix. 6) Repeat this until there have been a number of updates equivalent to 10% of the final epidemic size. We note that Lekone and Finkenstädt (2006) first utilized the 10% value and indicated that it balanced mixing and speed for their update algorithm. For our algorithm, we note

that updating 10% of the exposure yields very similar chains to a system where 50% of the exposures are updated, but runs faster.

This process varies the possible exposure times of each individual. Of course, one must find an appropriate place to start the MCMC chain, as it is not permissible for any individual to stay exposed for more than the maximum time, even in the starting condition of the chain. The prior information regarding the infectious disease is helpful here. We propose several chains be started at different possible values of the mixing and intervention parameters, and $\mathbf{E}$ generated via a simulation that makes use of the best guess for the latent period distribution. In our experience, this technique allows for the most variability in the starting conditions of the chains, so convergence is easy to assess.

2.2 *Properties of the Sampling Scheme*

The aforementioned algorithm gives good convergence for reasonable exposure distributions, such as gamma or Weibull distributions which are not exponentially distributed. It is important to note that without an intervention or with a constant intervention (outlined in Section 4 of this Web Appendix), convergence was attained on every attempt, but convergence failed with one particular data set using a small epidemic generated from exponential distribution times and an exponential decay intervention. This was because the algorithm would occasionally generate an exposure matrix where none of the individuals contracted a latent infection after the intervention began. The intervention parameter realizations would increase, and it became very difficult for the algorithm to return to realistic values. We do caution researchers when using this approach for exponentially or nearly exponentially distributed latent times to check convergence. If the exponential assumption holds, we recommend the Lekone and Finkenstädt model. When the exponential assumption is violated, our PS SEIR model has good convergence properties, and offers advantages over other models in the literature.

Additionally, with weak prior information and large data sets, the sampling algorithm

tends to overestimate the variance of the latent time distribution. This is acceptable, as the simulation results demonstrate that the algorithm is most helpful for small epidemics. In these cases, weak prior information easily contains this process. With large data sets, the algorithm can be still be used. The strong prior information typically available for the latent period can easily control this flattening of the conditional probabilities. If weak prior information is used, the algorithm will typically select a latent period where the conditional probabilities of compartmental change from the Exposed to Infectious compartments are more uniform than the prior information. This still yields advantages over the exponential assumption case, though these advantages are not as marked. When sampling unknown infectious paths via this scheme, where the total number moving into and out of the Infectious compartment are known at each time point, or if this data is available for the Exposed class, there is no flattening of the conditional probabilities of transferring from the Infectious to the Removed compartment, even with weak prior information and large data sets.

Additionally, when sampling from exponentially distributed latent and infectious time distributions, the MCMC chains of all the model parameters for the PS SEIR model draw from equivalent distributions to those of the population-averaged approach we outlined above. A proof of this claim can be found below. Web Table 1 demonstrates the equivalence between the MCMC chains generated by the population averaged and path-specific models. An epidemic consisting of 30 cases was simulated using exponential latent and infectious times. No intervention was used, and all infectious and removal times were assumed known. The quantiles provided were based on 7,000 iterations after burn-in. The mixing parameter posterior quantiles are almost exactly equal between the population averaged and path-specific model, but it may be noted that all the quantiles of the mean exposure time are slightly higher for the population averaged model than for the path-specific model. This is due to a practical consideration when using the PS SEIR model. The maximum time for the

latent period was set to be 100 days in the path-specific approach, removing approximately the longest 1% of latent times from consideration. Thus, the differences are due to the coding of the models, rather than a theoretical concern.

2.3 *A Special Case of the PS SEIR Model*

Given the assumption that we do not place a maximum limit on the amount of time that a patient spends in the latent category, we are able to demonstrate the equivalency of our model, as defined in Equation 2 of the main paper, with the population averaged model as defined in Equation 1 of the main paper. This will demonstrate that the Population Averaged model is a speical case of the PS SEIR model. The following assumptions are made:

1) Once in the exposed class, a patient must wait at least one time unit before moving to the infectious class.

2) Assume that the total number of patients entering and exiting the infectious class is known and fixed at every time point. This causes cancelation in the ratio of likelihoods for the intervention and mixing parameters when applying MCMC sampling. The results that hold for the exposed class also hold for the infectious class, with the only modification being that the number entering the class at any given time point is known.

3) Homogeneous mixing.

4) We consider a homogeneous population in terms of susceptibility.

5) There is only one individual in the infectious compartment, no individuals in the exposed compartment, and a fixed and known number in the removed compartment at the start of the epidemic.

6) We assume h is equivalent in both models. When this holds we assume, without loss of generality, h=1.

The following notation will be used: P_M will represent the full probability distribution of the population averaged model, and P_I for the PS SEIR model. In the population averaged

model, E_i will represent the total number of exposed individuals at time i, and E'_i will represent the number of individuals who are newly exposed on the i^{th} day of the epidemic. In the PS SEIR model, the notation $\mathbf{E}_{i,j}$ will be used, where each element will represent the number of individuals who have been exposed for $j * h$ days on the i^{th} day of the epidemic. For example, $\mathbf{E}_{i,1}$ represents the number of new exposures on day i. The notation $\mathbf{E}_{i+}$ will represent the marginal total of exposed individuals on day i. In both models, the same notation will be used for the infectious category, using I_i, I'_i, $\mathbf{I}_{i,j}$, $\mathbf{I}_{i+}$.

The full distribution for the population averaged model can be written as:

$$P_M(\underline{\psi}, \rho | E_i, i = 1, ..., T, I_i, i = 1, ..., T) = \prod_i L_{E'_i, S_i, \underline{\psi}} \binom{E_{i-1}}{I'_i} (1 - g(\rho))^{I'_i} g(\rho)^{(E_{i-1} - I'_i)} p(\underline{\psi}, \rho), \quad (1)$$

where i subscripts time, $\underline{\psi}$ represents all mixing and intervention parameters, and ρ represents the parameter of the exponential distribution for the exposed class, with $g(\rho) = \exp(-1/\rho)$. $L_{E'_i, S_i, \underline{\psi}}$ is the likelihood of the process producing new exposures. This process is assumed to have the same parameterization for both models.

In what follows, we show that an MCMC sampler will draw from the same posteriors for the parameters whether the specification is the population averaged model or the PS SEIR model.

By assumption, we know the number of new infectious individuals at day i. We assumed there was only one infectious individual at day 1, so S_i is fixed for all i. Note that with E_{i+} fixed and S_i known, we have $E_{i,1}$ fixed and known. Consider the PS SEIR model. For a given realization of $\mathbf{E}$, $E_{i+} = E_i$, where i subscripts the day of the epidemic, and j subscripts the bin of the exposed category.

The PS SEIR model can be written as:

$$P_I(\underline{\psi}, \rho, E_{i,j}) = \prod_i (L_{E_{i,1}, S_i, \underline{\psi}} \prod_j \binom{E_{i-1,j}}{E_{i,j+1}} (1 - g(\rho))^{I_{i1}} g(\rho)^{(E_{i-1,+} - I_{i,1})}) p(\underline{\psi}, \rho). \quad (2)$$

If we assume that $E_{i,1} = E'_i$ in the population averaged specification (and therefore E_{i+}=E_i) we have

$$P_M(\underline{\psi}, \rho | E_{i,1}, i = 1, ..., T, I_{i,1}, i = 1, ..., T) \propto P_I(\underline{\psi}, \rho | E'_i, i = 1, ..., T, I'_i, i = 1, ..., T).$$

Thus, for a given realization of the **E** in the PS SEIR model, the full distribution of the parameters in the PS SEIR model can be written in a population averaged form. Because the kernel of this full distribution of the PS SEIR model is proportional to that of the original population averaged model, it follows that the parameter realizations of the PS SEIR model come from the same posterior distribution as those from the original population averaged model whenever E_{i+}=E_i for all i.

To show that MCMC chains will draw from the same posterior distributions after burn in, it suffices to show that, given a realization of the parameters in the model, the distribution of E_{i1} and the distribution of E'_i are the same distribution. This follows because the full distributions for both the population averaged model and the PS SEIR model can be written in terms of $\underline{\psi}, \rho, E_i, E'_i$ under the assumptions in the population averaged model.

First, note that knowledge of the full set of infectious times $\{I'_i, i = 1, ..., T\}$ and initial conditions is sufficient to determine E_i from the set $\{E'_k, k \leqslant i\}$. Define Υ_i=$P(E'_i = E^{**}_i | E'_k = E^{**}_k \ \forall k < i, I'_k \ \forall k \leqslant i)$, under the population averaged model. Note that, given the starting conditions, the number of new infections, $I'_k \ \forall k \leqslant i$, and the number of new exposures ,E'_k $\forall k \leqslant i$, the total number exposed at time i, E_i is known. We denote this by E^*_i for ease of notation. Now given a realization of $\underline{\psi}, \rho$ and assuming that the total number of patients in the infective category is known at every time point, consider time $i|i-1$:

$$\prod_i \Upsilon_i = \prod_i (L_{E^{**}_i, S_i, \underline{\psi}} \binom{E^*_i}{I'_{i+1}} g(\rho)^{E^*_i - I'_{i+1}} (1 - g(\rho))^{I'_{i+1}}) p(\underline{\psi}, \rho). \tag{3}$$

Now, we turn to the PS SEIR model to show that $P(E_{i,1} = E^{**}_i | E_{k,1} = E^{**}_{k,1} \forall k < i, I_{k,1} \forall k \leqslant i) = \prod_i \Upsilon_i$. To begin, assume that all of the patients l=1,...,n in the PS SEIR model are distinguishable once exposed. The full posterior for the PS SEIR model can be written as the likelihood of $E_{i,1}$ newly exposed individuals, multiplied by a product of Bernoulli likelihoods, representing the other exposed bins. That is,

$$P_I(\underline{\psi}, \rho, E_{i,+,1}, E_{i,l,j\neq 1}) = \\ \prod_i (L_{E_{i,+,1},S_i,\underline{\psi}} \prod_l \prod_{j\neq 1} (1-g(\rho))^{I_{i,1,l}E_{i-1,l,j}} g(\rho)^{(E_{i-1,l,j}-I_{i,l,1}E_{i-1,l,j})}) p(\underline{\psi}, \rho) \tag{4}$$

where E_{ilj} is the indicator that person l is in exposed category j at time i, and $E_{i,+,1}$ is the number of new exposures at time i. For clarity on the terms $I_{i,l,1} * E_{i-1,l,j}$, consider the following: if person l is not in exposed category j at time $i-1$,then both $E_{i-1,l,j}$=0 and $I_{i,l,1} * E_{i-1,l,j}$=0 in the Bernoulli likelihood, and these terms do not change the value of full likelihood. If person l is in exposed category j at time $i-1$, then there are two options: 1) person l can become infectious, indicating the contribution to the likelihood is $(1-g(\rho))$ or 2) person l can remain exposed, with a contribution of $g(\rho)$, which are the conditional probabilities in the exponential assumption framework of 1) a person becoming infectious given they were exposed the previous day, or 2) staying exposed given they were exposed the previous day, respectively.

Note that knowledge of the set $\{E_{k,+,1}, k \leqslant i\}$ and the set $\{I_i\ , i = 1, ..., T\}$, along with the initial conditions (one individual in the infectious compartment, none in the exposed compartment, a fixed and known number in the removed compartment), fully determines E_{i++}, where $E_{k,+,l}$ is the number of individuals who entered the exposed class at time k, and E_{i++} is the population averaged total number of individuals in the exposed class at time i in the distinguishable framework.

Given a single path from $i-1$ to i, we have the following posterior:

$$P(E_{i,l,j\neq 1}, E_{i,+,1} | \underline{\psi}, \rho, E_{k,+,1} \text{ for all } k < i, E_{i-1,l,j}, I_{k,+,1} \text{ for all } k \leqslant i) = \\ L_{E_{i,+,1},S_i,\underline{\psi}} \prod_{j,l} (g(\rho)^{E_{ilj} - I_{i+1,l,1} * E_{ilj}} (1-g(\rho))^{I_{i+1,l,1}E_{ilj}} 1_{(E_{i-1,l,j-1} \geqslant E_{ilj})}) p(\underline{\psi}, \rho) \tag{5}$$

where we have assumed there is no maximum limit on the length of time a patient may stay in the exposed category. The indicator variables $1_{(E_{i-1,l,j-1} \geqslant E_{ilj})}$ are in place to indicate that paths through the exposed matrix are diagonally non-increasing. For example, it is

impossible for there to be an individual in exposed category j at time i, given that there was not an individual in exposed category $j-1$ at time $i-1$.

Next, we derive $P(E_{i,+,1} = E_i^{**} | I'_k \text{ for all } k \leqslant i, E_{k,+,1} = E_k^{**} \text{ for all } k < i)$ in the PS SEIR model. This will demonstrate that the distribution of the marginal sums for the PS SEIR model is equal to the probability distribution of the missing data for the population averaged model. To demonstrate this more clearly, we first consider a single path,

$$\begin{aligned} &P(E_{i,+,1} = E_i^{**} | I_{k,+,1} \text{ for all } k \leqslant i, E_{k,+,1} = E_k^{**} \text{ for all } k < i) = \\ &L_{E_{i,+,1}^{**},S_i,\underline{\psi}} \prod_{j \neq 1} \prod_l (g(\rho)^{E_{ilj} - I_{i+1,l,1} E_{ilj}} (1 - g(\rho))^{I_{i+1,l,1} E_{ilj}} 1_{(E_{i-1,l,j-1} \geqslant E_{ilj})}) p(\underline{\psi}, \rho), \end{aligned} \tag{6}$$

where we have evaluated the first product over l and considered only one particular path. The second product remains unevaluated because it is a fixed quantity for that given path.

To evaluate the second product, recall $E_{i++} = E_i^*$, which is fixed where the number of new exposures for $E_k, k \leqslant i$ and $I_k, k \leqslant i$, and the starting conditions of the system are known. Also note that $\sum_l \sum_j I_{i+1,l1} E_{ilj} = I_{i+1,+,1}$, which denotes all patients who were in an exposed class at time i and moved to the infectious class at time $i+1$. Thus, for a single path we write

$$\begin{aligned} &P(E_{i,+,1} = E_i^{**} | E_{k,+,l} = E_k^{**} \text{ for all } k < i, I_{k,+,1} \text{ for all } k \leqslant i) = \\ &\prod_i L_{E_{i,+,1}^{**},S_i,\underline{\psi}} g(\rho)^{E_i^* - I_{i+1,+,1}} (1 - g(\rho))^{I_{i+1,+,1}}. \end{aligned} \tag{7}$$

Now, consider the probability over all possible paths. We see the probability becomes

$$\begin{aligned} &P(E_{i,+,1} = E_i^{**} | E_{k,+,l} = E_k^{**} \text{ for all } k < i, I_{k,+,1} \text{ for all } k \leqslant i) = \\ &\prod_i L_{E_{i,1}^{**},S_i,\underline{\psi}} \binom{E_i^*}{I_{i+1,+,1}} g(\rho)^{E_i^* - I_{i+1,+,1}} (1 - g(\rho))^{I_{i+1,+,1}} = \prod_i \Upsilon_i. \end{aligned} \tag{8}$$

Summing across all valid paths yields the combinatorics. Considering only valid paths allows us to drop the indicator variables, as the cases relating to paths which are not valid yields a probability of zero. There are $\binom{E_i^*}{Ii+1,+,1}$ valid possible paths that yield the combinatoric, because, given an exposed vector at time i, we will have $I_{i+1,+,1}$ patients leaving the exposed class. Once these $I_{i+1,+,1}$ patients move out of the exposed class, the remaining patients will be forced to move diagonally through the exposure matrix. This demonstrates that the

distribution of the margins of the missing data in the PS SEIR model with distinguishable patients is the same as the distribution of the missing data in the population averaged model.

The previous result is for a given realization of patients. Now consider the patients to be indistinguishable. Note that when there are multiple patients in bin $E_{i,j}$ but fewer patients in bin $E_{i+1,j+1}$, this represents multiple potential paths. The number of paths represented is $\binom{E_{i,j}}{E_{i+1,j+1}}$. If we fix the values of the two rows of the exposure matrix (i.e. $E_{i-1,j}$ and $E_{i,j}$ are known for all j and a fixed i, then we see that the full posterior for $E_{i,1}$ given this particular arrangement of the exposure matrix is

$$\begin{aligned} &P(E_{i,1}|S_{k,+} \text{ for all } k \leqslant i, E_{k,1} \text{ for all } k \leqslant i, I_{k,+,1} \text{ for all } k \leqslant i) = \\ &\textstyle\prod_i L_{E_{i,1},S_i,\underline{\psi}} \prod_{j\neq 1}(\binom{E_{i-1,j-1}}{E_{i,j}} g(\rho)^{E_{i,j}-I_{i+1,1}E_{i,j}}(1-g(\rho))^{I_{i+1,1}E_{i,j}} 1_{(E_{i-1,j-1}\geqslant E_{i,+,j})}) p(\underline{\psi},\rho). \end{aligned} \tag{9}$$

This accounts for certain paths being more likely than others, but only accounts for the current arrangement of the exposure matrix. There will be many arrangements of the matrix that results from the same E_{i+1} totals.

Due to the relationship between the Bernoulli distribution and the binomial distribution, we know that summing across all possible arrangements with patients being indistinguishable will also yield a full probability $\prod_i \Upsilon_i$. This can be seen by a simple argument. If we consider all possible paths as arising from a set of Bernoulli distributions, and we know that there are $\binom{E_i^*}{I_{i+1,+,1}}$ possible combinations that will yield $I_{i+1,1}$ new exposures, we can view this set as summing to a binomial distribution. Alternatively, we can partition the Bernoulli random variables into groups, each as a binomial random variable with $E_{i-1,j-1}$ sample size and $E_{i,j}$ as the value taken by the random variable, as we have in the path-specific case. Now, the sum of all the Bernoulli random variables was a binomial as in the population averaged case (i.e. with $\binom{E_i^*}{I'_{i+1}}$ possible combinations), so we know that summing the binomial random variables arising from grouping the Bernoulli random variables will yield the same binomial random variable. This relies on the exponential assumption, because, for this proof to hold, the probabilities of success must be equal for all the bins, which is a consequence of the

exponential assumption. So, for the PS SEIR model with indistinguishable individuals, we have

$$\begin{aligned} &P(E_{i,1} = E_i^{**} | E_{k,1} = E_k^{**} \text{ for all } k < i, newI_k \text{ for all } k \leqslant i) = \\ &L_{E_{i,1}^{**}, S_i, \underline{\psi}} \binom{E_i^*}{newI_{i+1}} g(\rho)^{E_i^* - newI_{i+1}} (1 - g(\rho))^{newI_{i+1}} = \prod_i \Upsilon_i. \end{aligned} \tag{10}$$

This shows the distribution of E_{i1} in the PS SEIR model is equivalent to the distribution of E_i' in the population averaged model.

Therefore, we have proven that, given a realization of the missing data, the posterior distributions of the parameters are equivalent between the population averaged and path-specific models. We have also shown that, given a set of parameter realizations, the marginal distributions of the missing data are equivalent between the two models. Since the posterior distributions are known to be equivalent, the MCMC chains for both models will draw from the same posterior distributions after burn-in when the assumptions stated at the beginning of the proof hold.

This proof demonstrates that the PS SEIR model proposed in Equation 2 of the main paper is a generalization of the population averaged model found in Equation 1, and contains it as a special case.

3. Web Appendix C

For each simulated data set, there were 20,000 total individuals, with one member in the infectious category and all the other individuals susceptible at the start of the epidemic. The mixing parameter chosen to simulate the data was 0.25, in order to give a large degree of variability to the epidemic sizes. The form of the intervention was $f(\underline{\psi}, i) = \psi_1(1_{(i<i_0)} + (1 - \psi_2)1_{(i \geqslant i_0)})$. This represents a constant intervention, where the probability of moving from the susceptible to exposed class is decreased instantaneously at the time of the intervention, then held constant over the course of the intervention. For the simulations employing this intervention parameterization, ψ_2 was fixed at 0.7.

The parameterization selected for the exponential distributions was $f(x) = \frac{1}{\lambda}\exp(-\frac{x}{\lambda})$. For the Weibull distributions the parameterization was selected as $f(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}\exp(-(\frac{x}{\beta})^{\alpha})$. The true values for α and β in the Weibull distributions were chosen to be 7 and 20 for the exposed time, and 12 and 9 for the infectious time.

Table 2 presents the simulation results obtained for the constant intervention with Weibully distributed latent and infectious times, and all the removal times known. The PS SEIR model typically performs as well as, and often better than, the population averaged model in terms of the size of the credible interval and median parameter values. While the improvement offered by the PS SEIR model often appears to be small, it is important to note that small changes in parameter values can have a large effect on the distribution of final epidemic size in stochastic models. The priors used for the mixing and intervention parameters were Gamma(0.1,1) and Gamma(0.7,1) for both models. For the exponential analysis of each data set, λ was assigned a Gamma(187.09,10) prior. For the Weibull analyses, α was assigned a prior of Gamma(70,10) and β was assigned a Gamma(200,10) prior.

Table 3 presents simulation results demonstrating the difference in parameter posteriors using a constant intervention gamma distributed latent and infectious times when the removal times are known versus when they are imputed according to the correct distribution. When removal times are imputed to the correct distribution, parameter posteriors are very similar to those when the removal times are known.

References

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford Science Publications.

Lekone, P. E. and Finkenstädt, B. (2006). Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study. *Biometrics* **63,** 1170–1177.

Mode, C. J. and Sleeman, C. K. (2000). *Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases, and Computers*. World Scientific Publishing Co. Pte. Ltd.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

Table 1

Web Table 1: Equivalency of the Population Averaged Model and the Path-Specific Model.

Model	Parameter	2.5%	25%	50%	75%	97.5%
Population Averaged	Mixing	0 .088	0.113	0.132	0.151	0.185
Path-Specific	Mixing	0.088	0.114	0.131	0.149	0.185
Population Averaged	Mean Exposure Time	15.14	17.14	18.48	19.94	22.99
Path-Specific	Mean Exposure Time	14.98	17.01	18.36	19.81	22.73

Table 2

Web Table 2: Parameter medians and 95% central credible intervals for the analysis of the Weibull data sets featuring a constant intervention. Based on 7,000 realizations after burn-in each.

Analysis	Data Set	Mixing (0.25)	Intervention (0.7)	Alpha (7)	Beta (20)
Exponential Analysis	Small Weibull	0.23 (0.11, 0.38)	0.61 (0.14, 0.89)	Not Applicable	Not Applicable
Weibull Analysis	Small Weibull	0.25 (0.13, 0.39)	0.68 (0.25, 0.89)	6.86 (5.37, 8.51)	20.56 (18.08, 22.78)
Exponential Analysis	Medium Weibull	0.25 (0.16, 0.34)	0.70 (0.46, 0.86)	Not Applicable	Not Applicable
Weibull Analysis	Medium Weibull	0.26 (0.19, 0.34)	0.73 (0.59, 0.84)	6.39 (4.94, 8.04)	20.81 (18.23, 23.19)
Exponential Analysis	Large Weibull	0.26 (0.20, 0.33)	0.73 (0.59, 0.84)	Not Applicable	Not Applicable
Weibull Analysis	Large Weibull	0.24 (0.19, 0.30)	0.69 (0.56,0.79)	6.41 (5.03, 7.97)	19.06 (17.64, 20.61)
Exponential Analysis	Very Large Weibull	0.28 (0.22, 0.34)	0.79 (0.68,0.87)	Not Applicable	Not Applicable
Weibull Analysis	Very Large Weibull	0.26 (0.22, 0.31)	0.75 (0.67, 0.81)	6.33 (5.16, 7.83)	20.39 (19.27, 21.37)

Table 3

Web Table 3: Comparing parameter medians and 95% central credible intervals between known removal times and imputed removal times for a constant intervention with gamma latent and infectious time distributions

Imputation	Size	Mixing (0.25)	Intervention (0.7)	Alpha (30)	Beta (1.604)
Yes	Small	0.15 (0.08, 0.24)	0.30 (0.01, 0.66)	29.92 (26.71, 33.18)	1.75 (1.50, 2.02)
No	Small	0.15 (0.10, 0.24)	0.26 (0.01, 0.58)	29.97 (26.81, 33.16)	1.73 (1.48, 1.99)
Yes	Medium	0.22 (0.15, 0.31)	0.57 (0.35, 0.74)	29.92 (26.79, 33.31)	1.58 (1.36, 1.81)
No	Medium	0.23 (0.16, 0.31)	0.59 (0.35, 0.75)	30.11 (26.98, 33.51)	1.57 (1.36, 1.79)
Yes	Large	0.23 (0.18, 0.28)	0.68 (0.57, 0.77)	29.35 (26.24, 32.53)	1.71 (1.46, 1.99)
No	Large	0.22 (0.17, 0.27)	0.65 (0.49,0.76)	29.28 (26.12, 32.62)	1.72 (1.46, 1.99)
Yes	Very Large	0.25 (0.20, 0.30)	0.68 (0.59, 0.76)	29.03 (25.96, 32.36)	1.43 (1.24, 1.64)
No	Very Large	0.25 (0.20, 0.29)	0.67 (0.57, 0.76)	29.08 (25.90, 32.30)	1.44 (1.25, 1.63)