A frequentist approach to estimating the force of infection and the recovery rate for a respiratory disease among infants in coastal Kenya

Peer-reviewed author version

Mwambi, H.; Ramroop, S.; White, L.J.; Okiro, E.A.; Nokes, D.J.; SHKEDY, Ziv & MOLENBERGHS, Geert (2011) A frequentist approach to estimating the force of infection and the recovery rate for a respiratory disease among infants in coastal Kenya. In: STATISTICAL METHODS IN MEDICAL RESEARCH, 20(5), p. 551-570.

DOI: 10.1177/0962280208098666
Handle: http://hdl.handle.net/1942/10506

A frequentist approach to estimating the force of infection for a respiratory disease using repeated measurement data from a birth cohort

Mwambi H *, Ramroop S
University of KwaZulu-Natal,
P/Bag X01 Scottsville, PMB, South Africa
White LJ
Mahidol-Oxford Tropical Medcine Research Unit
420/6 Rotchawithi Rd., Bangkok 10400 Thailand
and
Centre for Tropical Medicine,
Nuffield Department of Clinical
Medicine, CCVTM, University of Oxford,
Churchill Hospital, Old Road,
Oxford, OX3 7LJ, United Kingdom.
Okiro EA, Nokes DJ
Kenya Medical Research Institute,
CGMRC, PO Box 230, Kilifi 80108 Kenya
Shkedy Z and Molenberghs G
Hasselt University, Agoralaan 1,
B-3590, Diepenbeek, Belgium

June 8, 2010

*email:MwambiH@ukzn.ac.za

Abstract

This paper aims to develop a probability based model involving the use of direct likelihood formulation and generalized linear modelling (GLM) approaches useful in estimating important disease parameters from longitudinal or repeated measurement data. The current application is based on infection with respiratory syncytial virus (RSV). The force of infection and the recovery rate or per capita loss of infection are the parameters of interest. However because of the limitation arising from the study design and subsequently the data generated only the force of infection is estimable. The problem of dealing with time varying disease parameters is also addressed in the paper by fitting piecewise constant parameters over time via the GLM approach. The current model formulation is based on that published in White et al. [27] with an application to rotavirus transmission and immunity.

Keywords: *Generalized Linear Models, Susceptible-Infected-Susceptible model, Recovery Rate, Respiratory Synctial Virus, Time-Dependent Force of Infection.*

1 Introduction

Respiratory syncytial virus (RSV) infection, which manifests primarily as bronchiolitis and/or pneumonia, is the leading cause of viral lower respiratory tract (LRT) infection in infants and young children. The clinical entity of bronchiolitis was described at least 100 years ago. In 1956, RSV, as the causative agent of most epidemic bronchiolitis cases, initially was isolated by Morris et al. [17] from chimpanzees with upper respiratory tract (URT) infections. Subsequently, Collins et al. [4] associated this agent with bronchiolitis and LRT infection in infants. Since then, multiple epidemiologic studies have confirmed the role of this virus as a leading cause of LRT infection in infants and young children. Cane and Pringle [1] states that human RSV causes LRT disease in about 40% of primary cases and is responsible for the hospitalization of $0.1\% - 2\%$ of infants under the age group of 1 year annually. Peak incidence of occurrence is observed at age 2–8 months. Overall, 3.5–4 million children younger than 4 years acquire an RSV infection, and in the United States alone more than 100,000 children are hospitalized annually because of this infection. This translates to 9–14 per 1000 children younger than 1 year who are hospitalized annually because of this condition. The virus does not induce solid immunity, re-infection is the norm (though progressively less severe), and, as yet no vaccine appears to be on the horizon. Virtually all children have had at least one RSV infection by their third birthday. Given the prevalence and potential severity of this condition, it is not surprising that the World Health Organization has targeted RSV for vaccine development. The frequency of RSV can be categorized as follows:

- Internationally: RSV infection is prevalent worldwide, with similar clinical manifestations and young age of RSV LRT infection;
- Race: All races appear susceptible to RSV, with similar disease patterns;
- Sex: Although boys and girls are affected equally by milder RSV disease, the frequency of hospitalization for RSV disease is higher in males, with a male : female ratio of approximately 2:1;
- Age: Severe RSV disease is primarily a disease of young infants and children, with a peak occurrence at age 2–8 months. Reinfection with RSV occurs throughout life, with disease becoming more limited to the URT.

In the field of infectious disease modelling, one area that is now re-attracting a lot of attention, is that of the statistical estimation of key parameters as-

sociated with disease processes. These key parameter estimates are based on observed data that is generated by the underlying disease process. In this paper we consider the estimation of the force of infection. It was not possible to estimate per capita loss of infection or recovery rate of the disease process. The disease of interest is a respiratory infection of children mainly under the age of one year. It is a viral disease caused by the Respiratory Syncytial Virus (RSV). Mathematical models to study the disease are not new. Greehalgh et al. [6] used both theoretical and deterministic models to study the RSV dynamics. Other relevant references on previous modelling work on RSV include Weber *et al.* [22] and White *et al.* ([25], [26]). In this paper we address the problem of combining the dynamics of the disease and the estimation of model parameters from observed data. The data used in our case are repeated measurements representing the status of whether a child is infected (1) or not (0) at a particular time point t_{ij} where the index i denotes an individual (child) and j denotes the observation occasion. Thus, we are faced with the problem of repeated non–normal data suggesting the use of statistical methods of analysis able to account for the correlation of responses within the same subject or cluster. In the current study we employ direct likelihood estimation and also discuss the use and implement the generalized linear modelling approach (McCullagh and Nelder, [11]) for the estimation of time varying stepwise force of infection and the per capita loss of infection (recovery rate). White *et al.* [27] solved a similar type of a problem using hierarchical Bayesian formulation to study rotavirus transmission and immunity. In this paper application to repeated measurements data was implemented via Markov Chain Monte-Carlo modelling using WinBuGs software. Thus the White *et al.* [27] method can also be applied to the current data set as an alternative method.

The description of the the Kilifi RSV Study and the available data are given in Section 2. In Section 3, the basic dynamics of RSV are discussed in relation to the SIS (Susceptible-Infected-Susceptible), SIR (Susceptible-Infected-Recovered) and SIRS models. In Section 4, we present how the estimation of the model parameters was carried out. One complicating factor in the process is that of the time varying disease parameters in the underlying process, hence the need to allow time dependence in the estimation of the parameters. A piecewise modelling approach was used to address this aspect. Section 5 is devoted to conclusions and suggestions of possible future extensions.

2 The Kilifi RSV Study

The Kilifi RSV study yields a repeated measurement (longitudinal) data set measuring the presence or absence of the Respiratory Syncytial Virus in children in coastal Kenya. A longitudinal study is one where data are obtained when a response is measured repeatedly on the same observational or experimental unit(s). The Kilifi data set is part of a study carried out by the Kenyan Medical Research Institute in collaboration with the Wellcome Trust in Kilifi, Kenya. The data used in this analysis comprise of a single birth cohort with observations primarily over the first year of life (Nokes *et al.*, [15]) and form part of a larger cohort study (Nokes *et al.*, [16]). The data set exhibits, simultaneously, several forms of so-called coarsening (Heitjan and Rubin [8], Zhang and Heitjan [29]), in the sense that the data structure assumed is richer than the data that are actually observed. First, the real underlying process of the disease is not directly observable, but only through the explicitly observed outcomes of the process. Second, the observations are not equally spaced within and between individuals and, importantly, the number of observations is not the same between individuals. This is less refined, or coarser, than the hypothetical observation of a continuous-time process. Third, information in between two observed events is unknown, additional events could have happened between any such pair of time points. Fourth, it is possible for children to drop out prior to the scheduled end of the study. This last form of coarsening is the more conventional missingness or dropout. A priori, it is possible for these coarsening processes, in particular dropout, to depend on (1) observed outcomes, (2) covariates and (3) unobserved (and unobservable) outcomes. If option (3) is the case, a so-called missing not at random (MNAR) mechanism is operating (Little and Rubin [9], Molenberghs and Kenward [12]) and, arguably, a wholly satisfactory analysis is beyond reach, and the most sensible route forward is by what is currently known as a sensitivity analysis, where a variety of complex models, accommodating MNAR, is considered. In this paper, we will make the assumption of missing at random (MAR), where missingness, or more generally coarsening, is allowed to depend on covariates and observed outcomes but, conditional upon these, not further on unobserved outcomes. This is considered by many a plausible assumption (for a review, see Molenberghs and Kenward [12]) and, very importantly, in a likelihood-based inferential framework, MAR is sufficient (provided some mild regularity conditions hold) to allow the analyst to ignore the missing data mechanism, i.e., there is no need to model it explicitly. In other words, one can proceed by fitting a model, such as a generalized linear model, using maximum likelihood, provided all data are subjected to analysis, from both completely and incompletely observed subjects. These

considerations imply that our analyses are valid under the assumption of MAR. Hence, this way of treating dropout is both broadly valid and, from a practical standpoint, does not require additional programming or otherwise technical work. Evidently, it might be of interest to conduct sensitivity analyses relative to the assumption of MAR, but this would go beyond the current research. The model that will be developed to represent this data will aid in understanding the process and in the design of more complex models in the future in order to be able to capture the kinds of incompleteness mentioned above. Proper inference about the disease process can be drawn through such models and eventually to aid in the design of intervention strategies. The Kilifi data set had 368 children that were recruited in the study, however only 334 childrens' data were measured and recorded. In total, there are 9374 responses that were measured and the number of times each child is measured varies from one child to another. For example, child number 344 is measured at 12 different occasions with unequal spaced time intervals while child number 368 was measured at 20 different occasions with equally spaced time intervals. Let Y_{ij} denote the outcome at observation time t_{ij} for individual i. Then assuming a first order Markov model (Diggle et al. [5]), the observed matrix of the number of transitions between the two states 'infected' and 'uninfected'can be represented as in Table 1 below. The table is for anyone who ever transited in the entire study period. It is meant to indicate the number of transitions into each of these states given the immediate past state. Thus they are conditional transitions. With reference to the current data set there are two important remarks about the state transitions. First the transitions from uninfected state to the infected state refer to symptomatic infections only since samples from children without at least mild symptoms or a cold were not collected. This means some infections will have been missed. Secondly following a confirmed RSV infection event it was assumed that the child was resistant to re-infection and hence no further sampling was scheduled for two weeks. This will clearly lead to an underestimate of 'infected state to infected state' transitions.

It should be noted that Table 1 gives the number of visits to the uninfected and infected states conditional on the previous state indicated by the row label. From the resulting matrix of transitions, it is clear that the rate of sampling far exceeded the rate of infection because most of the transitions were from uninfected to uninfected states. There are a total of 131 transitions among the children from the infected to the uninfected state. Similarly, there are 132 transitions from uninfected to infected states. This represents about 40% of infections in the first year of life. Furthermore the number

		Y_{ij}	
		uninfected	infected
Y_{ij-1}	uninfected	8598	132
	infected	131	13

Table 1: Matrix of the number of transitions into the infected and uninfected states conditional on the immediate past state

of transitions from infected to infected (Table 1) is small (only 13), given the high frequency of sampling, suggesting that the duration of infection is short (or equivalently high recovery rate). Later the assumption that the rate of recovery far exceeds the rate of infection is made. It is important to note that the time interval between transitions was not constant. The time intervals were different within and between the children which, as previously stated, makes the data set highly unbalanced. Therefore, standard methods of analysis may not be directly applicable.

3 The Model

In this section we discuss (for the purpose of analysis) the transmission of RSV for this particular cohort of children in relation to the SIS, SIR and SIRS disease models. In an SIS disease model, each individual in the population is either infected (I) or susceptible to infection (S). When a susceptible individual becomes infected, he/she is immediately infectious and when an infected individual is cured, he/she is immediately susceptible again. In a homogenous model assumption, every susceptible individual has the same probability of being infected, and each infected individual has the same probability of recovery. Ross [19] introduced the deterministic SIS model while Weiss and Dishon [23] introduced the stochastic SIS model namely as a Markov birth-and-death process that is used to model a variety of processes that range from epidemics, transmission of rumors and chemical reactions. It is also important to note that the long term behaviour of the deterministic and stochastic versions of the SIS model are quite different and we will not go into the details of this difference. In the current problem it should be noted that according to the biology of RSV the disease process may not necessarily follow an SIS disease model but rather a more appropriate model would be nearer to the SIRS process (White et al. [26]; Weber et al. [22]) with a

possibility of gradual immunity acquisition. RSV tends to occur in seasonal outbreaks, and while reinfections during one epidemic do occur, it tends to be the case that repeat infections occur in sequential epidemics.

However, in the first year of life there is little opportunity of reinfection since only one epidemic was experienced by the vast majority of this infant cohort, hence strictly speaking there is no basis for choosing between SIS, SIR or SIRS model. Further if the SIRS model framework is actually the more reasonable structure for RSV, and given the short period of follow up (i.e. little opportunity for loss of immunity and reinfection) then the more appropriate assumption would be to model the infection as a SIR structure.

We therefore restrict ourselves to modelling the process of primary infection and recovery, and we do this by using the simplest of forms where we model the transition rates from the disease free to the infected state (λ) and from the diseased state to the disease free state (ν) using an SIS type model. Thus according to the study design and data λ is correctly specified but ν is not and later it will be denoted by $\tilde{\nu}$ to distinguish it from the true value. The problem is to estimate the parameters of interest from observed data in the form of repeated (longitudinal) measures where each child presents a sequence of responses of 1's (diseased) and 0's (disease-free). The time duration between states (uninfected and infected) in days was also recorded thus the parameter estimates will have days^{-1} as units. We emphasize such estimates and their interpretations should always be carefully be linked to the study design and not from the data alone.

3.1 Model Governing Differential Equations

The SIS basic governing differential equation is given by

$$\frac{\partial q(a,t)}{\partial t} + \frac{\partial q(a,t)}{\partial a} = -\lambda(a,t)q(a,t) + \nu(a,t)p(a,t), \tag{1}$$

where $q(a,t)$ and $p(a,t)$ are, respectively, the proportion of susceptible and infected individuals in the population at time t and age a such that

$$p(a,t) + q(a,t) = 1.$$

Thus for a purely SIS model it is enough to study the solution for equation (1). However as already mentioned above RSV is a viral disease therefore the most appropriate model is the SIRS model where R is the class of recovered individuals with a possible loss of immunity to revert back to the S class. Thus in this case the equation for $p(a,t)$ would become

$$\frac{\partial p(a,t)}{\partial t} + \frac{\partial p(a,t)}{\partial a} = \lambda(a,t)q(a,t) - r(a,t)p(a,t)$$

where $r(a,t)$ is the rate at which individuals move from the infected state to the recovered class with a possible loss of immunity at rate $\nu^*(a,t)$ different from $\nu(a,t)$ in equation (1). But because the data currently in use was based on children within the age of one year the immunity against the disease for such individuals is still not yet developed therefore we assume $\nu^*(a,t) = 0$. It therefore suffices to deal with equation (1) ignoring the R to S transition as explained in the main opening paragraph under Section 3. For the sake of simplicity we also ignore the complication of short term immunity from infection in the first months of life due to maternally derived specific RSV antibodies. Hence, it is assumed that all children are born susceptible. In addition note that losses due to natural mortality can here be assumed to be balanced by new births therefore in effect we are assuming a constant population model. In the Kilifi data set all the children were all within one year of age, thus we can drop age, in equation (1) and therefore write

$$\frac{dq(t)}{dt} = -\lambda(t)q(t) + \nu(t)p(t). \tag{2}$$

If we assume $\lambda(t)$ and $\nu(t)$ are time-independent then

$$\frac{dq(t)}{dt} = -\lambda q + \nu(1-q) = -(\lambda+\nu)q + \nu \tag{3}$$

because $p(t) + q(t) = 1$. This equation can easily be solved using the 'variation of coefficients' technique (see Appendix A). Applying the technique to equation (3) a solution for $q(t)$ is obtained as:

$$q(t) = \frac{\nu}{\lambda+\nu} + \frac{\lambda}{\lambda+\nu}e^{-(\lambda+\nu)t}, \tag{4}$$

assuming $q(0) = 1$ and $p(0) = 0$ as the initial conditions and since $p(t)+q(t) = 1$ we get

$$p(t) = \frac{\lambda}{\lambda+\nu} - \frac{\lambda}{\lambda+\nu}e^{-(\lambda+\nu)t} \tag{5}$$

as the general solutions for $p(t)$. Note that if we relax the more restrictive initial condition that $q(0) = 1$ and $p(0) = 0$ and rather use the more general initial condition $p(0)+q(0) = 1$ the solutions for $p(t)$ and $q(t)$ are respectively given by

$$p(t) = \frac{\lambda}{\lambda+\nu} + \left(p(0) - \frac{\lambda}{\lambda+\nu}\right)e^{-(\lambda+\nu)t}$$

and

$$q(t) = \frac{\nu}{\lambda+\nu} + \left(q(0) - \frac{\nu}{\lambda+\nu}\right)e^{-(\lambda+\nu)t}$$

but for simplicity we stick to equations (4) and (5).

3.2 Linking the Model to Data

Note that the model solution for $q(t)$ implies that $q(\infty) = \frac{\nu}{\lambda+\nu}$ and hence $p(\infty) = \frac{\lambda}{\nu+\lambda}$ which give the equilibrium proportions of susceptible and infected individuals respectively. This means that for a rare disease we expect $\nu >> \lambda$. Now, let the indicators 1 and 0 denote respectively the infected and uninfected states of an individual and let Y_{it} denote a binary response variable taking on one of these values. The subscript i denotes a particular subject in the sample for $i = 1, \ldots, n$, where n is the number of subjects and t is time. Thus over a time interval $(0, t)$ we can define the four conditional state transition probabilities as follows:

$$\begin{aligned}
\pi_{00}(t) &= P(Y_{it} = 0 | Y_{i,0} = 0), \\
\pi_{01}(t) &= P(Y_{it} = 1 | Y_{i,0} = 0), \\
\pi_{10}(t) &= P(Y_{it} = 0 | Y_{i,0} = 1), \\
\pi_{11}(t) &= P(Y_{it} = 1 | Y_{i,0} = 1).
\end{aligned}$$

Suppose that at $t = 0$ the proportion infected is 0, that is $q(0) = 1$ and $p(0) = 0$. It is noted that since the disease process is reversible, individuals cannot remain infected forever. The solutions for $q(t)$ in (4) implies that, given an individual was initially uninfected, then the probability that this individual is still uninfected after a time duration t is given by,

$$\pi_{00} = \frac{\nu}{\lambda+\nu} + \frac{\lambda}{\lambda+\nu} e^{-(\lambda+\nu)t}, \tag{6}$$

and since $\pi_{01} + \pi_{00} = 1$, then

$$\pi_{01} = \frac{\lambda}{\lambda+\nu} - \frac{\lambda}{\lambda+\nu} e^{-(\lambda+\nu)t}. \tag{7}$$

Following similar arguments, we can write expressions for $\pi_{11}(t)$ and $\pi_{10}(t)$ as:

$$\pi_{11} = \frac{\lambda}{\lambda+\nu} + \frac{\nu}{\lambda+\nu} e^{-(\lambda+\nu)t}, \tag{8}$$

and

$$\pi_{10} = \frac{\nu}{\lambda+\nu} - \frac{\nu}{\lambda+\nu} e^{-(\lambda+\nu)t}. \tag{9}$$

Note that the process satisfies the ergodic property namely, $\pi_{00}(\infty) = \pi_{10}(\infty) = \frac{\nu}{\nu+\lambda}$ and $\pi_{01}(\infty) = \pi_{11}(\infty) = \frac{\lambda}{\nu+\lambda}$, the equilibrium proportion of susceptible and infected, respectively. Estimates of λ and ν can be obtained from these equations via maximum likelihood estimation since the transitions conditionally on previous state represent two separate Bernoulli distributions with

probabilities π_{01} and π_{10} or their complements, whenever necessary. The general form of the likelihood can be written as:

$$\left\{\prod_{i=1}^{N} P(Y_{i,0})\right\} \prod_{i=1}^{N} \prod_{j=1}^{n_i} P(Y_{i,j}|Y_{i,j-1}),$$

using the notation $Y_{i,j}$ to denote the binary observation from child i at time occasion j out of n_i occasions. The second part of the likelihood, obtained by conditioning on the previous measurement $Y_{i,j-1}$, is the same as that of a product of two independent Bernoulli likelihoods:

$$\prod_{i=1}^{N} \prod_{i=1}^{n_i} P(Y_{i,j}|Y_{i,j-1}) \propto (\pi_{01})^{n_{01}}(1-\pi_{01})^{n_{00}}(\pi_{10})^{n_{10}}(1-\pi_{10})^{n_{11}},$$

where $n_{k,l}$ are the total number of transitions from state $k \in (0,1)$ to state $l \in (0,1)$ and therefore explicit maximization is possible. There is an inherent assumption here that the time intervals are of equal length which in practice is not the case. It is possible to estimate the transition probabilities by maximizing this conditional likelihood instead of the full likelihood, since the initial measurement $Y_{i,0}$ contributes a limited amount of information only if some steady state assumptions are made. Thus, conditional on the initial states $\{Y_{i,0}\}$, the free parameters π_{01} and π_{10} are orthogonal. This allows a separate analysis of the two independent Bernoulli distributions leading to the maximum likelihood estimates of the two transition probabilities given

$$\hat{\pi}_{01} = \frac{n_{01}}{n_{01}+n_{00}}$$

and

$$\hat{\pi}_{10} = \frac{n_{10}}{n_{10}+n_{11}}.$$

Subsequently $\hat{\pi}_{00} = 1-\hat{\pi}_{01}$ and $\hat{\pi}_{11} = 1-\hat{\pi}_{10}$. Upon equating theses estimates of the transition probabilities to equations (7) and (9) (or equivalently working with their complements and equations (6) and (8)) one can ideally obtain estimates of the transition rates λ and ν. The problem with this approach is that the estimating equations are highly non-linear and the method works well for equally spaced observation times, as in Nagelkerke et al. [14]. In our case we are faced with a more complex situation. The observations are not equally spaced within and between subjects and in addition the number of observations is not constant over individuals. Thus we are dealing with a more complex scenario than that described in Nagelkerke et al. [14] requiring some simplifying assumptions. The alternative formulation adopted in Section 4 allows the use of generalized linear modelling approach. This

approach has an advantage of easily allowing for time varying (in our case monthly specific) parameters as will be seen in section 4.2 of the paper. We are not at all against the above approach but we are merely presenting an alternative approach to a similar problem.

4 Estimation of the Model Parameters

An alternative estimation procedure is developed by assuming that the residence or sojourn times in each disease state is exponentially distributed. As already explained the reason for changing to an alternative estimation procedure is that the data we are dealing with is highly unbalanced with unequal time intervals between sampling visits and in addition individuals do not all have equal number of observations. Thus we need some simplifying assumptions in order to easily work with the data via the generalized linear modelling (GLM) approach (sub-section 4.1). In the current model assume that the duration in the susceptible or disease free state is exponentially distributed with parameter λ. If recovery was possible then the duration in the disease state would be exponentially distributed with parameter r. Thus, we could correctly interpret λ and r as the force of infection and the recovery rate, respectively. The two parameters can also be viewed as the hazard of infection and recovery respectively. In effect we are assuming that the time of stay in the infected class is exponentially distributed with mean r^{-1} days. Likewise the time of stay in the susceptible class is assumed to be exponentially distributed with mean λ^{-1} days. Thus one can ideally consider two Poisson stochastic processes with exponential inter-arrival times. If we observe the processes within an interval of time $(0, d)$ we can infer that given an individual is in the susceptible class the probability of on infection at or before time d is $1-e^{-\lambda d}$ and the probability of no infection event is $e^{-\lambda d}$. Similarly given an individual is in the infected class the probability of a recovery at or before time d is $1 - e^{-rd}$ and the probability of no recovery event e^{-rd}. Thus conditional on the previous state we have two independent stochastic processes that need to be studied. This argument is the basis of the current formulation which was previously published by White *et al.* [27]. However after careful inspection of the full study design and the data generated, it became clear that it was not not possible to estimate the true recovery rate, r for RSV. Thus to emphasize this fact we change notation and use $\tilde{\nu}$ instead of r and $\bar{\nu}$ to denote an estimate of this parameter apparently estimable using the current data which should not be interpreted as the recovery rate. We therefore define the four observable transition probabilities for the current

data as follows:

$$\begin{array}{rcrcl} \pi_{00} & = & P(Y_{ij}=0|Y_{i,j-1}=0,d_{ij}) & = & e^{-\lambda d_{ij}}, \\ \pi_{01} & = & P(Y_{ij}=1|Y_{i,j-1}=0,d_{ij}) & = & 1-e^{-\lambda d_{ij}}, \\ \pi_{10} & = & P(Y_{ij}=0|Y_{i,j-1}=1,d_{ij}) & = & 1-e^{-\tilde{\nu} d_{ij}}, \\ \pi_{11} & = & P(Y_{ij}=1|Y_{i,j-1}=1,d_{ij}) & = & e^{-\tilde{\nu} d_{ij}}. \end{array} \tag{10}$$

The quantity $d_{ij} = t_{ij} - t_{ij-1}$, is the time interval between samples at time t_{ij} and t_{ij-1}. The full likelihood can therefore be written as:

$$L(\tilde{\nu},\lambda) = (P_0(1))^{\sum \delta_i}(P_0(0))^{N-\sum \delta_i} \prod_{0\to 0} e^{-\lambda d_{ij}} \prod_{0\to 1}(1-e^{-\lambda d_{ij}}) \prod_{1\to 0}(1-e^{-\tilde{\nu} d_{ij}}) \prod_{1\to 1} e^{-\tilde{\nu} d_{ij}}.$$

Now, δ_i is an indicator variable denoting the initial state of a child where $\delta_i = 1$ when the child is initially infected and 0 otherwise. Here $P_0(1)$ is the unconditional probability that the child is initially in the infected state. Likewise the unconditional probability that an individual is uninfected is $P_0(0) = 1 - P_0(1)$. If N is the total number of individuals in the study then $\sum \delta_i$ are individuals who are initially in the infected state. Since $P_0(1)$ and $P_0(0)$ are unknown it is simpler to consider the conditional likelihood given the initial states $Y_{i,0} \in \{P_0(1), P(0)\}$ in order to find the MLEs of the parameters λ and $\tilde{\nu}$.

Using the Fisher scoring method (see Appendix B) to iteratively solve for λ and $\tilde{\nu}$ the estimates together with approximate 95% confidence intervals are $\hat{\lambda} = 0.001169\,(0.000953, 0.001388)$ and $\bar{\nu} = 0.45495\,(0.32362, 0.58826)$ respectively. It should be noted that the estimate of the rate parameter $(\tilde{\nu})$ is high, compared to the estimate of the force of infection. We further emphasize that the time duration given by $\bar{\nu}^{-1}$ of 2 days cannot be interpreted as an estimate of the shedding duration of RSV based on the current data. The reason is because samples were not taken during infection. Thus the current data cannot support the estimation of the true recovery rate and hence the shedding duration. Based on observational studies carried out recently on the same population this duration is estimated to be between 4 and 11 days (Okiro et al. [18]). The current analysis is a very good example of a requirement in experimental design theory where it is stated that the analysis and therefore results of a designed study or experiment should directly be linked to the design. The estimated force of infection is justified and it is for infants in the primary phase of the disease where for simplicity we have assumed negligible maternal protection duration.

4.1 Application of GLM estimation to the RSV Data

As earlier defined, λ will denote the force of infection but $\tilde{\nu}$ will note denote the per capita loss of infection or the recovery rate for the disease process. If

we apply the generalized linear model to derive the force of infection for RSV, it will be necessary to consider data on the transitions from the uninfected to infected states namely, from state 0 to state 1 or $0 \rightarrow 1$ and the transitions from uninfected to uninfected that is $0 \rightarrow 0$. These transitions would make up 2 binary events for the response variable and once these transitions are coded as 1 for $0 \rightarrow 0$ and a 2 for $0 \rightarrow 1$, the response variable can be seen to conditionally follow a Bernoulli distribution. Likewise, another pair of binary responses can be similarly defined by considering the transitions $1 \rightarrow 1$ and $1 \rightarrow 0$. The residence times in the disease free and disease states are assumed to follow the exponential distribution with parameters λ and $\tilde{\nu}$, respectively. In survival analysis terminology, λ can also be interpreted as the hazard of infection or per capita risk of infection. The simpler model is where the only explanatory variable is the inter-state time duration that is, the quantity d_{ij}. Using generalized linear model (GLM) with log link function we obtain

$$\log(\pi_{00}) = -\lambda d_{ij}$$

and

$$\log(\pi_{11}) = -\tilde{\nu} d_{ij}$$

Since the data consist of 4 transition probabilities as defined in equation (10), in order to formulate an appropriate GLM we define an indicator variable

$$Z_{ij} = \begin{cases} 1 & Y_{ij} = 0, Y_{i,j-1} = 0, \\ 0 & Y_{ij} = 0, Y_{i,j-1} = 1, \\ 0 & Y_{ij} = 1, Y_{i,j-1} = 0, \\ 1 & Y_{ij} = 1, Y_{i,j-1} = 1. \end{cases}$$

Let $\theta_{ij} = P(Z_{ij} = 1)$ and consider the following linear predictor

$$\log(\theta_{ij}) = -\lambda d_{ij} \times (1 - Y_{ij}) - \tilde{\nu}_k d_{ij} \times (Y_{i,j-1}), \tag{11}$$

it follows that

$$\log(\theta_{ij}) = \begin{cases} -\lambda d_{ij} & \text{if } Y_{ij} = 0, Y_{i,j-1} = 0, \\ -\tilde{\nu} d_{ij} & \text{if } Y_{ij} = 1, Y_{i,j-1} = 1. \end{cases} \tag{12}$$

Thus, using this approach we obtained $\hat{\lambda} = 0.0021$ (95% C.I: 0.0018-0.0024) and $\bar{\nu} = 0.503$ (95% C.I: 0.386-0.657) for the force of infection and the parameter ν, respectively. Again as with estimates found using direct likelihood maximization the force of infection leads to a disease free duration of about 1.5 years and the estimate $\bar{\nu}$ leads to a duration of 2 days which as earlier stated cannot be interpreted as the shedding duration of RSV for this population of infants (Okiro et al. [18]).

4.2 Time Dependent Force of Infection

The above estimation procedures only helped us to estimate a single constant force of infection and per capita loss of infection over the time period of the study. However, there is enough evidence that a disease such as RSV does exhibit clear temporal variation in its incidences, which is a function of the force of infection. Thus, we extended the above approach to obtain monthly piecewise estimates of the force of infection. For months 14 and 15, there are no data because none of the children completed the study up to months 14 and 15 hence no estimate is available for these two months. A piecewise constant force of infection with log link function was assumed. Hence, the linear predictor is given by

$$\log(\theta_{ij}) = -\lambda_k d_{ij} \times (1 - Y_{ij}) - \tilde{\nu}_k d_{ij} \times (Y_{i,j-1}). \tag{13}$$

Here, λ_k is the monthly force of infection but we re-emphasize that $\tilde{\nu}_k$ is not the per capita loss of infection or recovery rate. Note that the model in (13) can be re-expressed in terms of a complementary-log-log link, in which the linear predictor is given by

$$g(\theta_{ij}) = -\log(\lambda_k) \times (1 - Y_{ij}) - \log(\tilde{\nu}_k) \times (Y_{i,j-1}) + \log(d_{ij}) \tag{14}$$

where g is the complementary-log-log link function. In such a model, the monthly regression parameter estimates for the force of infection and the parameter ν are equal to $\log(\lambda_k)$ and $\log(\tilde{\nu}_k)$, respectively. As a result, the parameter estimates for the monthly force of infection and the additional parameter $\tilde{\nu}$ are constrained to be non- negative, as required. In this paper, the complementary-log-log link function was used to estimate the model's parameters. 95% confidence intervals were obtained either by exponentiating the model parameters and their confidence intervals or by applying the delta method for the log of the parameters. Table 2 presents the parameter estimates for the monthly force of infection. The force of infection peaks with different heights in months 3 ($\hat{\lambda}_3 = 0.007$), then it decreases to zero at month 9 and increase to secondary peaks at months 11 and 12 ($\hat{\lambda}_{11} = 0.0022$ and $\hat{\lambda}_{12} = 0.0022$, respectively). Month 1 had too few transitions recorded in it while months 14 and 15 did not have any data in them since the children did not complete the study for these months. Hence, these months have been omitted in the analysis. Figure 1 shows a plot of the force infection against time together with 95% confidence intervals from direct exponentiation and the delta method. There is virtually no difference between the two sets of confidence intervals.

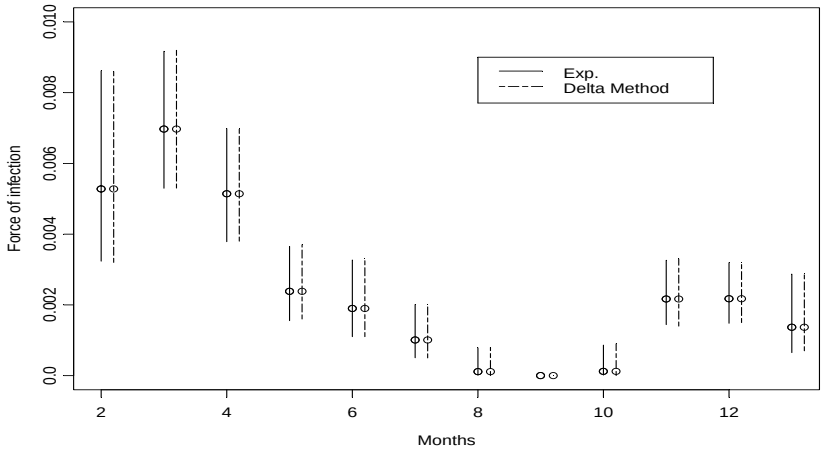

Figure 1: The force of infection in months together with 95% confidence intervals using the exponentiated and delta methods.

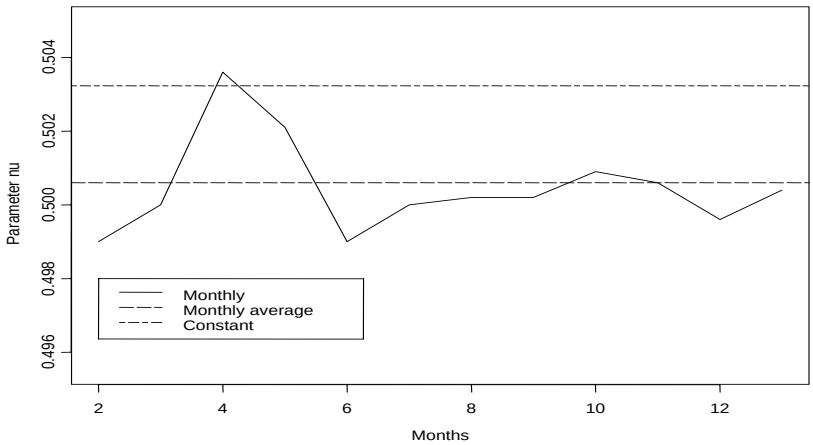

Figure 2: A plot of bar nu in months.

		Exponentiation		Delta Method		
Month	Lambda	Estimate (day^{-1})	95% C. I.		95% C. I.	
2	$\hat{\lambda}_2$	0.0053	0.0032	0.0086	0.0027	0.0079
3	$\hat{\lambda}_3$	0.0070	0.0053	0.0092	0.0051	0.0089
4	$\hat{\lambda}_4$	0.0051	0.0038	0.0070	0.0036	0.0067
5	$\hat{\lambda}_5$	0.0024	0.0016	0.0037	0.0014	0.0034
6	$\hat{\lambda}_6$	0.0019	0.0011	0.0033	0.0009	0.0029
7	$\hat{\lambda}_7$	0.0010	0.0005	0.0020	0.0003	0.0017
8	$\hat{\lambda}_8$	0.0001	0.0000	0.0008	-0.0001	0.0003
9	$\hat{\lambda}_9$	0.0000	0.0000	0.0000	0.0000	0.0000
10	$\hat{\lambda}_{10}$	0.0001	0.0000	0.0009	-0.0001	0.0004
11	$\hat{\lambda}_{11}$	0.0022	0.0014	0.0033	0.0013	0.0031
12	$\hat{\lambda}_{12}$	0.0022	0.0015	0.0032	0.0013	0.0030
13	$\hat{\lambda}_{13}$	0.0014	0.0007	0.0029	0.0004	0.0024

Table 2: Monthly estimates of the force of infection and confidence Intervals

For completeness the monthly estimates of the parameter $\tilde{\nu}$ were also similarly obtained and the values are tabulated below for comparison purposes. However we re-iterate that these do not represents estimates of monthly recovery rate of RSV.

Month	Nu	Estimate(day^{-1})	Standard Error
2	$\bar{\nu}_2$	0.4990	0.067
3	$\bar{\nu}_3$	0.5000	0.06
4	$\bar{\nu}_4$	0.5036	0.064
5	$\bar{\nu}_5$	0.5021	0.062
6	$\bar{\nu}_6$	0.4990	0.066
7	$\bar{\nu}_7$	0.500	0.076
8	$\bar{\nu}_8$	0.5002	0.072
9	$\bar{\nu}_9$	0.5022	0.065
10	$\bar{\nu}_{10}$	0.5009	0.06
11	$\bar{\nu}_{11}$	0.5006	0.071
12	$\bar{\nu}_{12}$	0.4996	0.061
13	$\bar{\nu}_{13}$	0.5004	0.069

Table 3: Monthly estimates of the per capita loss of infection

Months 14 and 15 did not have any data in them because none of the children completed the study up to months 14 and 15. The estimate of the parameter $\tilde{\nu}$ is fairly constant over all the months with no unusual peaks in the estimates. Graphically the monthly estimates of $\tilde{\nu}$ are plotted over the study period as shown in Figure 2. Since from Table 3 we see that the monthly estimates of $\tilde{\nu}$ were very stable within a very narrow range we opted for a common recovery rate estimate. A formal likelihood ratio test was performed to compare the two models (constant versus monthly specific estimates) and the difference was not statistically significant (LR statistic =11.36 on 11 d.f and p-value=0.4743). Thus a combined estimate was calculated by finding a weighted average of the 12 estimates and the variance of this common estimate calculated by weighting the within and between component variances (see Appendix C). However because of the similarity of the 12 monthly specific values the contribution from the between component variance was very small and negligible. The overall estimate of the transition rate $\tilde{\nu}$ estimated this way was $\bar{\nu}_1 = 0.5006$ with a SE of 0.0189 and an approximate 95% CI given by $(0.4635, 0.5377)$. The two horizontal lines in Figure 2 represent the combined estimate by the method above (lower horizontal line, $\bar{\nu} = 0.5006$) and the common rate parameter $\tilde{\nu}$ from a GLM (upper horizontal line, $\bar{\nu}_2 = 0.5032$). A Wald test for a difference between these two estimates shows they are not significantly different.

5 Conclusion

In this paper generalised linear modelling combined with likelihood estimation was used to estimate the force of infection for a childhood respiratory disease (RSV). In the process an additional parameter $\tilde{\nu}$ associated with the data was also estimated. Estimation using the full likelihood was not possible therefore a form of conditional likelihood was used to model the data. The generalised modelling approach was modified to estimate monthly specific force of infection for the disease thus allowing the model to capture the temporal trends of disease incidence via piecewise parameter estimation. The force of infection was estimated as $\hat{\lambda} = 0.0012$ and the rate parameter $\tilde{\nu}$ is estimated as $\bar{\nu} = 0.4550$ using the direct maximum likelihood estimation method. Corresponding estimates using the generalized linear modelling approach are 0.0021 and 0.5032. These two approaches gave quite similar sets of parameter estimates for the parameter $\tilde{\nu}$ but the GLM approach yielded a force of infection around twofold higher. However we prefer the GLM appraoch because of its flexibility in allowing us to come up with monthly piecewise parameter estimates. It is also seen from the estimation of the

monthly parameters that RSV force of infection peaks at month 3, 11 and 12 which correspond to the months of May, January and February according to the original study period. This is consistent with the discussions by Cane [2], Chew et al. [3] and Simoes [21] who all state that RSV has a seasonal signal attributed to meteorological or sociological factors. Furthermore, the force of infection is not constant and varies with time. It will be important at this point to discuss the validity of the estimates in relation to the limitations of the data collected in the study. First we argue that the force of infection is from both methods an underestimation not because the methods are wrong but because of failure in the study design to collect data on asymptomatic infections. Secondly the parameter $\tilde{\nu}$ by both methods cannot be used to derive the shedding duration because samples were not taken over the two week period following infection. The issue of average shedding duration has recently been reviewed by Okiro *et al.* [18]. Other shedding studies show that the viral load starts declining only after 4 days or so (Hall *et al.*, [7]). In summary it should be noted that a data generation process is a reflection of the study design which should be linked to the analysis and results. The parameter estimates also imply that the equilibrium proportion of susceptible and infected children stabilizes at around 99.74% and 0.26% largely the result of very short duration of infection. Note that from this analysis the susceptible prevalence is an estimate of both naive individuals and those treated from a previous infection and re-entered the S class. Nonetheless, since the force of infection is actually quite high, a significant proportion of infants are infected in the first year of life where disease risk and severity are highest. Thus, statistical and mathematical models are an important tool in understanding its dynamics and hence assist in designing control and intervention strategies for it. Further analyses to investigate child to child heterogenous effects and to account for the different forms of incompleteness mentioned in Section 2 are currently in progress. A sensitivity analysis to assess the impact of different forms of missing data types on the stability of parameter estimates is also proposed.

Acknowledgements

The authors gratefully acknowledge the financial support from The Wellcome Trust (grant No. 061584), and the IUAP research network Nr. P5/24 of the Belgian Government (Belgian Science Policy). Shaun Ramroop would like to thank the NRF of South Africa for funding his PhD work (THUTHUKA-Researchers in training Ref. No: TTK2005081700004). Mahidol-Oxford Tropical medicine Research Unit is funded by the Wellcome Trust of Great

Britain. The paper is published with the permission of the Director of KEMRI.

References

[1] Cane P, Pringle C. Molecular epidemiology of respiratory syncytial (RSV) virus: rapid identification of subgroup A isolates. *Journal of Virological Methods* 1992; **40**: 297–306.

[2] Cane, P. Molecular and epidemiology of respiratory syncytial virus. *Reviews in Medical Virology* 2001; **11**: 103–116.

[3] Chew F, Doraisingham S, Ling A, Kumarsinghe G' Lee B. Seasonal trends of viral respiratory tract infections in the tropics. *Epidemiology and Infection* 1998; **121**: 121–128.

[4] Collins PL, McIntosh K, Channock RM. Respiratory syncytial virus, In Fields BN, Knipe DM, Howely PM, eds. *Fields Virology*, Lippincott-Raven, Phliadelphia; 1996. pp. 1313–1351.

[5] Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*, 2nd edn. Oxford University Press, Great Britain; 2002.

[6] Greenhalgh D, Deikmann O, de Jong MCM. Subcritical endemic steady states in mathematical models for animal infections with incomplete immunity. *Mathematical Biosciences* 2000; **165**: 1–25.

[7] Hall C, Douglas RJ, Geiman J. Respiratory syncytial virus infections in infants: quantitation and duration of shedding. *Journal of Pediatrics* 1976; **89**: 11–15

[8] Heitjan DF, Rubin DB. Ignorability and coarse data. *The Annals of Statistics* 1991; **19**: 2244–2253.

[9] Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York; 2002.

[10] Longford NT. (1993) *Random Coefficient Models*, 2nd ed. Oxford University Press, New York; 1993.

[11] McCullagh P, Nelder JA. *Generalised Linear Models*, 2nd edn. London: Chapman and Hall, London; 1989.

[12] Molenberghs G, Kenward MG. *Missing Data in Clinical Studies.* John Wiley & Sons, Chichester; 2007.

[13] Molenberghs G, Verbeke G. *Discrete Models for Longitudinal Data*, 1st edn. Springer Verlag, New York; 2005.

[14] Nagelkereke G, Chunge RN, Kinoti SN. Estimation of parasite infection dynamics when dectectability is imperfect. *Statistics in Medicine* 1990; **9**: 1211–1219.

[15] Nokes DJ, Okiro E, Ngama MJ, White LJ, Ochola R, Scott PD, Cane PA, Medley GF. RSV epidemiology in a birth cohort in Kilifi District, Kenya: infection in the first year of life. *Journal of Infectious Diseases* 2004; **190**: 1828–1832.

[16] Nokes DJ, Okiro EA, Ngama MJ, et al. (2008). Respiratory syncytial virus infection and disease in infants and young children observed from birth in Kilifi District, Kenya. *Clin Infect Dis* 2008; **46**: 50–57.

[17] Morris JA, Blount RE, Savage RE. Recovery of cytopathogenic agent from chimpanzees with coryza *Proceedings of the Society of Experimental Biology and Medicine* 1956; **92**: 544–549.

[18] Okiro EA, White LJ, Ngama M, Cane PA, Medley GF, Nokes DJ.(2010). Duration of shedding of respiratory syncytial virus in a community study of Kenyan children. *BMC Infect Dis.* 2010; Jan **22**: 10–15.

[19] Ross, R. Some a priori pathometric equations. *British Medical Journal* 1915; **1**: 546.

[20] Searle SR, Casella G, McCulloch CE. *Variance Components*. John Wiley & Sons, Inc, New York; 1992.

[21] Simoes, E. (1999). Respiratory syncytial virus infection. *Lancet* **354**, 847–852.

[22] Weber A, Weber M, Milligan P. Modeling epidemics caused by respiratory syncytial virus (RSV). *Mathematical Biosciences* 2001; **172**: 95–113.

[23] Weiss GH, Dishon M. On the asymptotic behaviour of the stochastic and deterministic models of an epidemic. *Mathematical Biosciences* 1971; **11**: 261–265.

[24] Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal data*, 1st edn. Springer Verlag, New York; 2000.

[25] White LJ, Waris M, Cane PA, Nokes DJ, Medley GF. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *Epidemiol Infect* 2005; **133**: 279–289.

[26] White LJ, Mandl JN, Gomes MG, Bodley-Tickell AT, Cane PA, Perez-Brena P, Aguilar JC, Siqueira MM, Portes SA, Straliotto SM, Waris M, Nokes DJ, Medley GF. Understanding the transmission dynamics of respiratory syncytial virus using multiple time series and nested models. *Math Biosci* 2007; **209**: 222–239.

[27] White LJ, Buttery J, Cooper B, Nokes DJ, Medley GF. Rotavirus within day care centres in Oxfordshire, UK: characterization of partial immunity. *J R Soc Interface* 2008; **5**: 1481–1490.

[28] RSV information website http:www.rsvinfo.com

[29] Zhang J, Heitjan DF. Impact of nonignorable coarsening on Bayesian inference. *Biostatistics* 2007; **8**: 722–743.

Appendix A: Solutions for $p(t)$ and $q(t)$

The steps to the solution of the SIS governing differential equation (3) are outlined below. Put the linear equation in the standard form as

$$\frac{dy}{dt} + G(t)y = f(t).$$

The integrating factor of the standard form is given by $e^{\int G(t)dt}$. Next multiply the standard form of the equation by the integrating factor and note that the left hand side of the resulting equation is automatically the derivative of the product of the integrating factor and y that is,

$$\frac{d}{dt}[e^{\int G(t)dt}y] = e^{\int G(t)dt}f(t).$$

Lastly integrate both sides of this last equation and solve for y subject to the initial conditions of the system.

Thus, the solution to the equation

$$\frac{dq(t)}{dt} = -\lambda q + \nu(1-q),$$

can be constructed by first noting that,

$$\begin{aligned} \frac{dq(t)}{dt} &= -\lambda q + \nu - \nu q, \\ \Rightarrow \frac{dq(t)}{dt} &= -(\lambda+\nu)q + \nu, \end{aligned}$$

implying that

$$\frac{dq(t)}{dt} + (\lambda+\nu)q = \nu.$$

Multiplying both sides by the integrating factor yields

$$\begin{aligned} e^{(\lambda+\nu)t}\frac{dq(t)}{dt} + e^{(\lambda+\nu)t}(\lambda+\nu)q &= e^{(\lambda+\nu)t}\nu \\ \frac{d}{dt}[e^{(\lambda+\nu)t}q(t)] &= \nu e^{(\lambda+\nu)t} \\ \int \frac{d}{dt}[e^{(\lambda+\nu)t}q(t)] &= \int \nu e^{(\lambda+\nu)t}dt \\ e^{(\lambda+\nu)t}q(t) &= \frac{\nu}{\lambda+\nu}e^{(\lambda+\nu)t} + c \\ q(t) &= \frac{\nu}{\lambda+\nu} + ce^{-(\lambda+\nu)t} \end{aligned}$$

Imposing the initial condition $q(0) = 1$ and $p(0) = 0$, that at $t = 0$ the proportion infected is 0 implies that we can solve for c as $c = 1 - \frac{\nu}{\lambda+\nu} = \frac{\lambda}{\lambda+\nu}$. Thus we can find $q(t)$ and $p(t)$ by using the condition $p(t) + q(t) = 1$.

Appendix B: Fisher scoring method equations

The full likelihood can be written as:

$$L(\nu,\lambda,dt) = (P_0(1))^{\sum \delta_i}(P_0(0))^{N-\sum \delta_i} \prod_{0\to 0} e^{-\lambda d_{ij}} \prod_{0\to 1}(1-e^{-\lambda d_{ij}}) \prod_{1\to 0}(1-e^{-\nu d_{ij}}) \prod_{1\to 1} e^{-\nu d_{ij}}$$

where δ_i is an indicator variable denoting the initial state of a child with $\delta_i = 1$ when the child is initially infected and 0 otherwise. Hence $P_0(1)$ is the probability that the child is initially in the infected state such that $P_0(0) = 1 - P_0(1)$, N is the total number of individuals in the study and $\sum \delta_i$ are the individuals who are initially in the infected state and $N - \sum \delta_i$ are initially not infected. It is thus simpler to consider the conditional likelihood given the initial states $\{Y_{i,0}\}$ in order to find the maximum likelihood estimates (MLEs) of the parameters λ and ν. If we take the log of the likelihood we get the log-likelihood as:

$$\ell = \log L = \log(constant) - \lambda \sum_{0\to 0} d_{ij} + \sum_{0\to 1} \log(1-e^{-\lambda d_{ij}}) + \sum_{1\to 0} \log(1-e^{-\nu d_{ij}}) - \nu \sum_{1\to 1} d_{ij}$$

Taking the first and second partial derivative with respect to λ and ν gives us the following set of equations to work with:

$$\begin{aligned}
\frac{\partial \ell}{\partial \lambda} &= -\sum_{0\to 0} d_{ij} + \sum_{0\to 1} [1/(1-e^{-\lambda d_{ij}})](e^{-\lambda d_{ij}})(d_{ij}) \\
\frac{\partial \ell}{\partial \nu} &= -\sum_{1\to 1} d_{ij} + \sum_{1\to 0} [1/(1-e^{-\nu d_{ij}})](e^{-\nu dij})(d_{ij}) \\
\frac{\partial^2 \ell}{\partial \lambda^2} &= -\sum_{0\to 1} [1/(1-e^{-\lambda d_{ij}})]^2 [(e^{-\lambda d_{ij}})d_{ij}]^2 + \sum_{0\to 1} [1/(1-e^{-\lambda d_{ij}})](e^{-\lambda d_{ij}})d_{ij}^2 \\
\frac{\partial^2 \ell}{\partial \nu^2} &= -\sum_{1\to 0} [1/(1-e^{-\nu d_{ij}})]^2 [(e^{-\nu d_{ij}})d_{ij}]^2 + \sum_{1\to 0} [1/(1-e^{-\nu d_{ij}})](e^{-\nu d_{ij}})d_{ij}^2 \\
\frac{\partial^2 \ell}{\partial \nu \partial \lambda} &= \frac{\partial^2 \ell}{\partial \lambda \partial \nu} = 0
\end{aligned}$$

Next the Fisher's scoring method to iteratively solve for λ and ν is briefly described. Longford [10] and Searle et al. [20] state that the Fisher's scoring method is preferred to Newton-Raphson method since it avoids the heavy computational burden of finding the Hessian matrix (the matrix of second derivatives of the log-likelihood) by using the inverse of the information matrix I^{-1} (i.e. replace the Hessian by the negative of its expected value, which is often easier to compute than the Hessian). The inverse of the information matrix will be required to get the estimated variance-covariance matrix of

our parameters. For generality purposes let the parameters λ and ν to be contained in a vector θ. The iterative scheme is then given by:

$$\theta^{(m+1)} = \theta^{(m)} + [I(\theta)^{(m)}]^{-1} \left[\frac{\partial \ell}{\partial \theta}\right]_{\theta=\theta^{(m)}}$$

where the superscript (m) denotes the m^{th} iteration and $I(\theta)^{(m)}$ is the estimate of the information matrix given $\theta = \theta^{(m)}$. The process is repeated until convergence.

Appendix C: Common recovery rate

Assume we have M ($M = 12$ in our case) estimates $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_M$ and that each has an estimated variance $V_1, V_2, \cdots, V_M$ respectively, calculated as the square of its SE. For each estimate we create a weight given by

$$W_i = \frac{\frac{1}{V_i}}{\sum_{i=1}^{M} \frac{1}{V_i}}$$

for $i = 1, 2, \cdots, M$. Then we calculate our overall common recovery rate (PLOI) as a weighted average $\tilde{\theta}$ as:

$$\tilde{\theta} = \sum_{i=1}^{M} W_i \, \hat{\theta}_i$$

whose variance is given by

$$V = \sum_{i=1}^{M} W_i^2 \, V_i + \sum_{i=1}^{M} W_i^2 (\hat{\theta}_i - \tilde{\theta})^2$$

and standard error given by $\text{SE}_V = \sqrt{V}$.