Title: An approach for modeling cross-immunity of two strains, with application to variants of Bartonella in terms of genetic similarity

Authors: Kwang Woo Ahn[a,*], Michael Kosoy[b], and Kung-Sik Chan[c]

Affiliation: [a]Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI USA

[b]Centers for Disease Control and Prevention, Fort

Keywords: *Bartonella*, conditional least squares, cross-immunity, SIR model

1. Introduction

Multi-strain models have been widely used in epidemiology (Gupta, Ferguson, and Anderson, 1998; Kamo and Sasaki, 2002; Abu-Raddad et al, 2005; Bianco et al, 2009; Minayev and Ferguson, 2009). Developing and using multi-strain models is a challenging procedure due to numerous parameters such as death rate, birth rate, force of infection, and transmission rate, which are commonly assumed to be strain specific.

One of the key concepts of these models is cross-immunity, which allows infection by one strain to induce partial/perfect protection against other strains. Gupta, Ferguson, and Anderson (1998) proposed a very general model accounting for the cross-immunity in a multi-strain system, based on which they studied the effects of cross-immunity on evolution of strain structure. Abu-Raddad and Ferguson (2005) investigated population dynamics of host-pathogen systems involving an arbitrary number of antigenically distinct strains whose interaction depends on the cross-immunity. Minayev and Ferguson (2009) studied multi-strain deterministic epidemic models in which cross-immunity varies with the genetic distance between strains. Kamo and Sasaki (2002) proposed a two-strain susceptible-infected-recovered (SIR) model with cross-immunity. These models, however, assume an equilibrium population size over time, i.e., equal, constant birth and death rates. These assumptions might be too strong since the host population might fluctuate dramatically between seasons, which may affect the force of infection (Davis et al, 2005). For example, hispid cotton rat populations usually have peak litter production occurring in late spring and in late summer-early fall (Cameron and Spencer, 1984).

In addition, these earlier works were restricted to simulation studies assuming known parameter values. Another issue is that, except for the model of

County, Georgia, USA, over a period of 17 months, from March, 1996 to July, 1997, except December 1996, yielding altogether 483 trapping records (Kosoy et al., 2004a; Kosoy et al., 2004b). Cotton rats were captured for two or three consecutive nights each month and blood samples were taken. First-time captured cotton rats were marked. Marked and sampled rats were released. Sixty four out of 483 trapped rats were found to have co-infections by two or three *Bartonella* strains. Based on the

strain and the prevalence of strain B was low. In this paper, we consider two scenarios of cross-immunity: i) between genogroups; ii) between variants in the same genogroup. For the first scenario, we combined B and C due to low frequency of strain B and relatively high genetic similarity between strains B and C (Table 1 of Kosoy et al., 2004b). For the second scenario, we consider genogroup A only because of its high prevalence. Table 1 of Kosoy et al. (2004b) shows that A1 and A5 are genetically close to each other, and so are A2 and A4. Therefore, in this report, we compare A1&A5 vs. A2&A4 and A vs. B&C.

3. A two-strain SIR model with state variables expressed as proportions

We consider the two-strain special case of the multi-strain model proposed by Gupta, Ferguson, and Anderson (1998). Their model provides a general framework for modeling the dynamics of an infectious disease with multiple strains of a pathogen that may induce various degrees of cross-immunity in the hosts. Here, we extend their model to allow for variable host reproduction, and that the death rate can also be variable and not equal to the birth rate. Moreover, we modify the model so that a host is assumed to only make a fixed number of contacts with other hosts, on

respect to the second strain. For example, x_{IS} is the proportion of hosts infected by the first strain but susceptible to the second strain, $x_{I\,x}$ is the proportion of hosts infected by the first strain, while x_{x} is the proportion of hosts infected by the second strain. All state variables are implicit functions of time with their derivatives denoted by the dot notation. The extended two-strain SIR model is given as follows:

$$
\begin{aligned}
x &= {}_1xy_1 + {}_2xy_2 + (1 - x)\mu, \\
y_1 &= {}_1(x + z_2)y_1 - ({}_1 + b)y_1, \\
y_2 &= {}_2(x + z_1)y_2 - ({}_2 + b)y_2, \\
z_1 &= {}_1xy_1 - {}_2{}_1y_2 + bz_1, \\
z_2 &= {}_2xy_2 - {}_1{}_2y_1 + bz_2,
\end{aligned}
\tag{1}
$$

where the parameter β_i's are the transmission rates between an individual infected by strain i (i=1, 2) and one susceptible to both strains, γ_i's are the host's recovery rate from an infection by

where due to a weakened immune system by on-going infection, a host infected by one strain and susceptible to the other strain may have an elevated chance of being infected by the latter strain compared to a host susceptible to both strains (Small et al., 2010). For such cases, C may be greater than 1. Thus, we shall allow C to be non-negative. In summary, $G = 0$ represents the case of perfect cross-immunity between the two strains. If C is positive and less than 1, there exists a partial cross-immunity between the two strains. If C is equal to 1, there is no cross-immunity for the two strains and they infect the host independently. For $G > 1$, it signifies that the two strains are positively correlated, i.e., infection by one strain elevates the transmission rate of the other strain to the host. Asymmetric cross-immunity (Nuño et al., 2008) may be incorporated into the above model. However, in view of the relatively shortness of the *Bartonella* data, a parsimonious modelirate (-)Tj 0.002 Tc -6( )-10(a)433 0 Td [(i)-10(mmu)2(n)2(i

The main scientific question we address here concerns how similarity between two bacterial strains as deduced from their genetic sequences may relate to the host's cross-immunity to the strains. We explore this issue by estimating the host's cross-immunity against two pathogen groups of strains using the dataset discussed in Section 2. We first fit the proposed two-strain SIR model (1) to the monthly Bartonella infection rates by A1&A5 vs. A2&A4, where subgroups are combined due to their low prevalence and their relatively close genetic similarity. We then repeat the analysis contrasting A vs. B&C, with B and C merged into a group for a similar reason. In each analysis, the monthly observations consist of proportions of caught hosts infected by each of the two strains under study. Specifically, let the infection rate of the sampled hosts in the t^{th} month by strain be denoted by $y_{i,t}, t \quad 1,2$. These observed infection rates differ from the population infection rates $y_{i,t}$ by an additive measurement error: $y_{i,t} \quad y_{i,t} \quad H_{i,t}$, where

$$b \quad b_t \quad u\sin(\frac{}{6}t) \quad v\cos(\frac{}{6}t) \quad w,$$

where *u*, *v* and *w* are unknown parameters.

The method of (approximate) conditional least squares via the unscented Kalman filter (UKF-CLS) (Ahn and Chan, 2013a) was employed to analyze the differential equation model (1) with the Bartonella data, which we now briefly outline. Consider the case that the state vector of the underlying system evolves according to a vector differential equation, with observations of some function of the state vector taken over discrete time. In our model, $y_t \quad (y_{1,t}, y_{2,t})^T$ is the observation vector and $v_t \quad (x_t, y_{1,t}, y_{2,t}, z_{1,t}, z_{2,t})^T$ the true state vector at time *t*. (For ease of exposition, we assume data were taken over equally-spaced epochs, say, $t \quad 1, 2, ..., n,$ but the method can be readily extended to irregularly sampled data.) Were the differential equation (1) linear and assuming normally distributed measurement errors, Kalman filter (an

approximation of the conditional means and variances for nonlinear processes, see Ahn and Chan (2013b). Unknown parameters can be estimated via approximate conditional least squares by minimizing the objective function $\sum_{t=1}^{n} | y_t - \hat{y}_{t|t-1}(\theta) |^2$

only if information of individual trapped host is available, which is, fortunately, the case for the *Bartonella* data.

5. Results

Tables 1 and 2 summarize the fitting results of the proposed two-strain SIR model with symmetric cross-immunity and identical recovery rates to the infection time series data with A1&A5 vs. A2&A4 and those with A vs. B&C, respectively. All 95% confidence intervals are obtained by nonparametric bootstrap detailed at the end of Section 4. The parameter $\mathcal{L}$

1.164 – 15.438) for the two-strain model A1&A5 vs. A2&A4, and that of A, vs. B&C, respectively. These results are consistent with the observations that *Bartonella* infections by these strains were endemic, with infections predominantly by strain A, in the cotton-rat population under study.

Note that the estimates of the birth rate parameters *u*, *v*, and *w* are similar in both models. The bottom left curves in Fig. 2 show the estimated birth rate curve. The curve suggests that birth rate attains the maximum in June and the minimum in December, which is consistent with the report by Rose (1986) reporting that all of the trapped female cotton rats were pregnant from March through July, but none were pregnant in November and December. Recall that $\hat{y}_{t|t\ 1}(\ \ \eta)$ is the approximate conditional mean of y_t, given past monthly observations, computed via the UKF. The fitted values in the t^{th} month are then given by $\hat{y}_{t|t\ 1}(\ \hat{\ }\eta)$ where $\hat{\ }\eta$ is the UKF-CLS estimate; the two components of the vector of fitted values will be denoted by $\hat{y}_{i,t}, t \quad 1,2.$ The fitted values (red X's) are joined by red solid lines in Fig. 2, superimposed with the 95% predictive bounds (blue dotted lines) of the infection rates in Fig. 2, which track the observed infected proportions (solid circles) well. The 95% predictive intervals are computed by the formula $\hat{y}_{i,t} \quad 1.96\sigma\ s_{i,t}$ where $s_{i,t}$ is the square root of the corresponding diagonal element of $Var_{\hat{\ }\eta}(y_{(t|t\ 1)})$ which is computed via UKF.

The residuals are defined as $r_{i,t} \quad y_{i,t} \quad \hat{y}_{i,t}$, i.e., subtracting the conditional means (fitted values) from the observed values, and the residuals estimate the error terms $e_{i,t} \quad y_{i,t} \quad E_{\hat{\eta}}(y_{(t|t\ 1)})$ where

γ_0 is the true parameter. By construction, the $e_{i,t}$'s are independent. So, the goodness of fit of the fitted models may be assessed by checking whether the residuals are approximately independent. Residual diagnostics may be further simplified by standardizing the residuals by normalizing them by the estimate of the conditional standard deviations computed via the UKF. We examine whether or not the (standardized) residuals are autocorrelated by checking the residual autocorrelation functions (ACF), while between-series dependence in the residuals can be examined by the cross-correlation function (CCF), which are plotted in Fig. 3. None of the residual autocorrelations are significant and so are all cross-correlations, except for one lag, suggesting that the standardized residuals are

3. Ahn, K. W., Chan, K. S. 2013b. On the convergence rate of the unscented transformation. To appear in Ann. I. Stat. Math.
4. Bianco, S., Shaw, L. B., Schwartz, I. B. 2009. Epidemics with multistrain interactions: The interplay between cross immunity and antibody-dependent enhancement. Chaos 19: 043123.
5. Cameron, G. N., Spencer, S. R. 1984. Sigmodon hispidus. Mammalian Species 158, 1–9.
6. Chan, K. S., Kosoy, M. 2010. Analysis of multi-strain Bartonella pathogens in natural host population – Do they behave as species or minor genetic variants? Epidemics 2, 165-172.
7. Clark, D. O. 1972. The extending of cotton rat range in California – Their life history and control. Proceedings of the 5th Vertebrate Pest Conference.
8. Davis, S, Calvet, E, Leirs, H. 2005. Fluctuating Rodent Populations and Risk to Humans from Rodent-Borne Zoonoses. Vector Borne Zoonotic Dis. 5, 305-314.
9. Gupta, S., Ferguson, N., Anderson, R. M. 1998. Chaos, persistence and the evolution of strain structure in populations of antigenically variable infectious agents. Science 240, 912-915.
10. Julier, S. J., Uhlmann, J. K. 1977. A new extension of the Kalman filter to nonlinear systems. Proc. of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Orlando, FL Vol. Multi Sensor Fusion, Tracking and Resource Management II.
11. Kamo, M., Sasaki, A. 2002. The effect of cross-immunity and seasonal forcing in a multi-strain epidemic model. Physica D 165, 228–241.

12. Kalman, R. E. 1960. New approach to linear filtering and prediction problems. J. Basic Eng-T ASMA D. 82, 35-45.
13. Kosoy, M., Hayman, D. T. S., Chan, K. S. 2012. Bartonella bacteria in nature: Where does population variability end a species start? Infect. Genet. Evol. 12, 894-904.
14. Kosoy, M., Mandel, E., Green, D., Marston, E., Jones, D., Childs, J. 2004a. Prospective studies of Bartonella of rodents. Part I. Demographic and temporal patterns in population dynamics. Vector Borne Zoonotic Dis. 4, 285–295.
15. Kosoy, M., Mandel, E., Green, D., Marston, E., Jones, D., Childs, J. 2004b. Prospective studies of Bartonella of rodents. Part II. Diverse infections in a single community. Vector Borne Zoonotic Dis. 4, 296–305.
16. Kosoy, M., Regnery, R.L., Kosaya, O.I., Childs, J.E., 1999. Experimental infection of cotton rats with three naturally occurring Bartonella species. J. Wildlife Dis. 35, 275–284.
17. La Scola, B., Zeaiter, Z., Khamis, A., Raoult, D. 2003. Gene-sequence-based criteria for species definition in bacteriology: the Bartonella paradigm. Trends Microbiol. 11, 318-321.
18. Minayevlds 3 >>BDC -29.92 -2.3(i)nI4(oul)-7(i)nds nergn . M 2093. e64(r52(e64ae64l(is1(m (o-2fr52n d

20. Nuño, M, Feng, Z., Martcheva, M., and Castillo-Chavez, C. 2005. Dynamics of two-strain influenza with isolation and partial cross-immunity, SIAM J. Appl. Math. 65, 964-982.

21. Rose, R. K. 1986. Reproductive strategies of meadow voles, hispid cotton rats, and eastern harvest mice in Virginia. Va. J. Sci. 37, 230-239.

22. Small, C. L., Shaler, C. R., McCormick, S., Jeyanathan, M., Damjanovic, D., Brown, E. G., Arck, P, Jordana, M., Kaushic, C., Ashkar, A. A., Xing, Z. 2010. Influenza infection leads to increased susceptibility to subsequent bacterial superinfection by impairing NK cell responses in the lung. J. Immunol. 184, 2048-2056.

Tables

Table 1. Parameter estimates of model (3) fitted to A1&A5 vs. A2&A4

	D_1	D_2	J	C
Estimates	1.931	1.722	0.079	0.132
95% CI	(1.			

Figure 1. Prevalence of *Bartonella* strains and the number of trapped cotton rats over the study period. Solid circles show observed values. The y-axis of the bottom figure represents the number of trapped cotton rats.

Figure 2. Estimated birth rate, prevalence and 95% confidence intervals. For the birth rate curve, the solid line is the birth rate function from the fitted model with A1&A5 vs. A2&A4 and the dotted line from that with A vs. B&C. The two estimated curves are quite similar. For the prevalence curves, the black dots and the red solid lines represent the observed and predicted infection rates, respectively. The blue dotted lines are the 95% confidence intervals.

Figure 3. Diagnostics: ACF, CCF, and Ljung-Box p-value plots. All estimated residual autocorrelations lie within the (individual) 95% confidence intervals (dotted lines), suggesting that the residuals are not auto-correlated. The estimated residual cross-correlations are also within the 95% confidence intervals, indicating no cross-correlations in the residuals. All p-values in Ljung-Box plots are greater than 0.05 (dotted line), which further confirms that the residuals appear to be white noise.

categories are denoted by X_{SS}, X_{IS}, etc., all of which are functions of time *t*, although the

Kamo and Sasaki (2002) assumed that σ varies between 0 and 1. For $\sigma = 0$, a host recovered from an infection by one strain of the pathogen acquires perfect cross-immunity against other strain, that is, it will never be infected by the other strain. On the other hand, for $\sigma = 1$, a host recovered from an infection by one strain of the pathogen does not have a cross-immunity against the other strain, so that the two strains can independently infect a host. One of the assumptions for the two-strain SIR model (1) is that a host may make contact with any host in the population and that the contact rate is proportional to the population size, which may not hold for the case of vast study area and/or large population.

To derive the two-strain model with the state expressed in terms of proportions, we consider the first derivative of the proportions, for example, $\frac{d}{dt}(X_{SS} / N)$. The variables standing for the proportions will be denoted in lower case, e.g. we write x_{SS} for X_{SS} / N, etc. Following Kamo and Sasaki (2002), we transform the nine state variables into five state variables as follows: $x = x_{SS}$, $y_1 = x_{IS}$, $y_2 = x_{RS}$, $z_1 = x_{SI}$, $z_2 = x_{SR}$, where $x_{IS} = x_{IS} + x_{II} + x_{IR}$ and $x_{RS} = x_{SI} + x_{II} + x_{RI}$. In addition, we replace the term $\beta_1 X_{SS} X_{IS}$ by $\beta_1 X_{SS} X_{IS} / N$ where β_1 is the product of the expected number of contacts a host makes with other hosts per unit time and the transmission probability given a contact between an individual infected by strain 1 and one susceptible to both strains. This specification implies that on average a host makes a fixed number of contacts per unit of time.

Next, we consider the first derivative of X_{SS}, $(X_{SS} / N)'$. Then, we have

$$\frac{X_{SS}}{N}\left(\frac{X_{SS}N \cdot X_{SS}N}{N^2} - \frac{X_{SS}}{N} - \frac{X_{SS}N}{N^2}\right)$$

$$\frac{{}_1X_{SS}X_I + E_2X_{SS}X_{xI} - X_{SS} - bNE}{N} - \frac{X_{SS}(bN - N)}{N^2}$$

$$\frac{{}_1X_{SS}X_I + E_2X_{SS}X_{xI} - bX_{SS} - bNE}{N}$$

$${}_1x_{SS}x_I + D_2x_{SS}x_{xI} - (1 - x_{SS})bD$$

Note that F

$$
\begin{aligned}
x' &= -\beta_1 x y_1 - \beta_2 x y_2 + (1 - x)\mu, \\
y_1' &= \beta_1 (x - \delta z_2) y_1 - (\gamma_1 + \mu) y_1, \\
y_2' &= \beta_2 (x - \delta z_1) y_2 - (\gamma_2 + \mu) y_2, \\
z_1' &= \delta \beta_1 x y_1 + \delta \beta_2 z_1 y_2 - \mu z_1, \\
z_2' &= \delta \beta_2 x y_2 + \delta \beta_1 z_2 y_1 - \mu z_2,
\end{aligned}
\tag{2}
$$

where the parameter β_i's are the transmission rates between an individual infected by strain i (i=1, 2) and one susceptible to both strains. Note that the death rate parameter μ is eliminated in the algebra so that it no longer appears in (2). However, since Equation (2) is directly derived from (1), (2) accounts for the possibility that the birth rate differs from the