Statistics and Its Interface Volume 12 (2019) 11–19

A latent spatial piecewise exponential model for interval-censored disease surveillance data with time-varying covariates and misclassification

Yaxuan Sun*, Chong Wang*,†,‡, William Q. Meeker*, Max Morris*, Marisa L. Rotolo†, and Jeffery Zimmerman†

Understanding the dynamics of disease spread is critical to achieving effective animal disease surveillance. A major challenge in modeling disease spread is the fact that the true disease status cannot be known with certainty due to the imperfect diagnostic sensitivity and specificity of the tests used to generate the disease surveillance data. Other challenges in modeling such data include interval censoring, relating disease spread to distance between units, and incorporating time-varying covariates, which are the unobserved disease statuses. We propose a latent spatial piecewise exponential model (PEX) with misclassification of events to address the challenges in modeling such disease surveillance data. Specifically, a piecewise exponential model is used to describe the latent disease process, with spatial distance and time-varying covariates incorporated for disease spread. The observed surveillance data with imperfect diagnostic tests are then modeled using a binary misclassification process given the latent disease statuses from the PEX model. Model parameters are estimated through a Bayesian approach utilizing non-informative priors. A simulation study is performed to evaluate the model performance and the results are compared with a candidate model where no misclassification is considered. For further illustration, we discuss an application of this model to a porcine reproductive and respiratory syndrome virus (PRRSV) surveillance data collected from commercial swine farms.

Keywords and phrases: Bayesian, Disease surveillance, Interval-censored data, Misclassification, Piecewise exponential model, Spatial.

1. INTRODUCTION

Animal pathogen control is a serious production, economic, and, in the case of zoonotic agents, a public health issue. Effective disease surveillance is key to achieving disease control, but understanding the dynamics of disease spread is critical to achieving effective animal disease surveillance. Therefore, the purpose of this paper is to use disease surveillance data to study disease transmission among pens within barns and determine the impact of distances among pens on disease spread.

Porcine reproductive and respiratory syndrome virus (PRRSV) surveillance data was used in this study because of its severe impact on pig production throughout the world. First identified on the basis of clinical signs in the 1980s, PRRSV had become endemic in most swine producing countries by the mid-1990s and is currently the most costly infectious disease of swine in many parts of the world. For example, PRRSV costs the United States swine industry approximately $664 million annually (Holtkamp et al., 2013). In China, PRRSV outbreaks caused pork prices to increase by 85 percent in 2006 (Lin et al., 2013). A common cause of respiratory and reproductive disease in pigs, PRRSV produces a chronic, persistent infection and stimulates weak protective immunity against genetically heterologous isolates (Zimmerman et al., 2012). These features present a severe challenge to control of the virus.

Animal disease surveillance data typically consists of diagnostic test outcomes for samples repeatedly collected at regular time intervals. This represents several challenges to understanding disease spread. First, the diagnostic sensitivity and specificity of the assays used to test samples are typically imperfect, so the true disease status is not known with certainty. Second, the latent time-to-disease is interval-censored because the samples are taken at predetermined time points, thus even discounting the uncertainty of the diagnostic test, the time-to-disease can only be known to have occurred within certain time intervals. Third, to study the spread of infectious agents, the hazard of a certain pen becoming infected needs to be modeled by incorporating the true disease statuses of other pens within the same barn, as well as the spatial distances among the pens. In this situation, the true disease statuses of the pens are time-varying covariates in the survival model and cannot be known with certainty.

Interval-censored data is commonly used in time-to-event analysis and abundant work has been done on the estimation

*Department of Statistics, College of Liberal Arts and Science, Iowa State University.
†Department of Veterinary Diagnostic and Production Animal Medicine, College of Veterinary Medicine, Iowa State University.
‡Corresponding author.

of survival functions in this situation. The study of interval-censored time-to-event data can be found in many areas, including clinical trials where patients underwent disease assessment at prescheduled clinic visits, animal epidemiology studies where samples were collected and tested at fixed sampling points, and other longitudinal studies (Lindsey and Ryan, 1998; Huang and Wellner, 1997; Finkelstein, 1986). In the present study, if the diagnostic test outcome perfectly reveals the true disease status at sequential time points, the interval-censored time-to-event can be determined to lie in the interval between the last negative test and the first positive test. However, with imperfect diagnostic test, the exact time interval where the event occurred is unknown, so the traditional estimation methods can no longer be applied directly. In recent years, the effect of misclassification and measurement error has received much attention. Lyles et al. (2011) used validation data-based adjustment for outcome misclassification in logistic regression. Yi et al. (2015) derived methods for mixed measurement error and misclassification in covariates. For survival models, McKeown and Jewell (2010) proposed a nonparametric maximum likelihood approach for misclassified univariate current status data. García-Zattera, Hara, and Komárek (2016) proposed an accelerated failure time model for misclassified clustered interval-censored data.

In this paper, for each pen, we evaluate the influence of the infection status of other pens in the same barn due to disease spread. The hazard rate of each pen can vary across the sampling periods. In addition, the spatial distances among pens must be included in the model so that the analysis can be used to determine whether, and to what extent, distance plays a role in disease spread. In this case, a piecewise constant hazard function can be used to account for the change of hazard; such a survival model is called piecewise exponential model. A piecewise exponential model is commonly used for interval-censored time-to-event data (Friedman, 1982; Lindsey and Ryan, 1998), where a constant hazard can be assumed in each time interval and covariate effects can be accommodated using proportional hazards, if there are any. Due to its simplicity and flexibility, this model has been advocated in different research areas in recent years. Berry et al. (2004) built a piecewise baseline hazard function in the Bayesian survival model to allow for changes in hazard rate overtime. Moreover, the piecewise constant hazard assumption was proved useful in some medical research areas, such as cancer survival analysis (Goodman, Lib, and Tiwari, 2011). However, the cited examples did not consider misclassification of the event outcome.

In this paper, we extend the piecewise exponential model to account for misclassification of the disease status as a result of the use of imperfect diagnostic tests. With predetermined jump points, we parameterize the log-hazard function for each pen with a piecewise linear regression. To account for disease spread, the covariates are set as the unknown true disease status for all pens within the same barn and the effect of their spatial distance is included and assessed in model. Sensitivity and specificity of the diagnostic test are used to build a model relating true disease statuses and the imperfect test outcomes.

This paper is structured as follows. In Section 2, the data structure and misclassification process are described, then the piecewise constant hazard is constructed for the unobserved time-to-event using a log-link and the latent disease statuses are incorporated as covariate effects. We present the likelihood function of the model and use a Bayesian approach for estimation of model parameters. In Section 3, the proposed model is applied to a PRRSV surveillance data set to illustrate the efficiency of the model. In Section 4, some simulation studies are performed to evaluate the proposed model. The estimation bias, standard deviation and RMSE (root mean square error) of the model parameters are reported from estimation results. Model without misclassification is run and compared to the proposed model. In Section 5, some conclusions and future work are discussed.

2. THE MODEL

Suppose there are I groups (buildings) and J subjects (pens) in each group $i(i = 1, 2, ..., I)$ in the study. Each subject $j(j = 1, 2, ..., J)$ is sampled at predetermined sampling time points $0 = \tau_0 < \tau_1 < ... < \tau_K < \infty$. Let u_{ijk} be the observed binary diagnostic test outcome for subject j in group i at sampling point τ_k. Test outcome u_{ijk} equals 1 if it is test positive and 0 otherwise. Let y_{ijk} denote the corresponding binary true event (disease) status, which equals 1 if the true status is diseased and 0 otherwise. The time origin is set to 0 and p_0 is introduced to be the probability that an event has happened at the enrollment of the study, i.e., $p_0 = p(y_{ij0} = 1)$. The diseased subjects are assumed to be diseased till the end of the study, i.e., if $y_{ijk} = 1$, then $y_{ijm} = 1$ for all $m \geq k$. For subjects that are free of an event (i.e., $y_{ij0} = 0$) at time 0, let t_{ij} denote the unobserved time-to-event for subject j in group i. Note that the true status y_{ijk} is uniquely determined by t_{ij}, i.e., $y_{ijk} = I(t_{ij} \leq \tau_k)$ for $k \geq 1$. In our model, animals once infected are assumed to be diseased and infectious throughout the study period. This is reasonable for the application to the PRRSV data in this paper as animals were never treated and remained positive once infected (refer to Section 4 for real data analyses). Consequently, there is no recurrent process.

To account for the spread of disease among subjects, distance between subjects needs to be defined. Figure 1 illustrates the internal structure of a barn (group) and the pen (subject) locations within the barn on a commercial swine farm. The subjects can be treated as arranged in a row of unit squares and each vertex of the square indicates a subject location. The number of pens within a barn may vary between production systems, but the inside structure of the barns are usually similar. Let $d_{jj'}$ denote the spatial distance between subjects j and j' within each group. If subjects j and j' are adjacent to each other in the unit grid then

Figure 1. Inside structure of a barn and the pen locations on a swine farm.

$d_{jj'} = 1$; otherwise, $d_{jj'}$ can be calculated as the Euclidean distance between j and $j^{'}$. Let $\mathbf{D}_i$ be a $J \times J$ distance matrix of $d_{jj'}$ for group i. For an animal disease surveillance data where the barns have equal size and structure, the $\mathbf{D}_i$ are the same for all $i = 1, 2, ..., I$. We are interested in modeling the time-to-event distribution for the subjects with consideration of the spatial spread of the infection among the subjects. We propose to describe the underlining time-to-event with a piecewise exponential model and relate the observed test outcome with the latent disease status using a misclassification model.

2.1 The misclassification model

Let γ_1 and γ_0 be the sensitivity and specificity of the diagnostic test. Sensitivity, also called the true positive rate, measures the proportion of actually diseased subjects which are correctly identified as positive. Specificity, also called the true negative rate, measures the proportion of actually healthy subjects which are correctly identified as negative. Based on the notations above, we have $\gamma_1 = p(u_{ijk} = 1|y_{ijk} = 1)$ and $\gamma_0 = p(u_{ijk} = 0|y_{ijk} = 0)$. Then for each subject j in group i at sampling point τ_k, the distribution of the diagnostic test outcome u_{ijk} given the latent disease status y_{ijk} can be defined as:

$$\begin{aligned} u_{ijk}|y_{ijk} = 0 &\sim \text{Bernoulli}(1 - \gamma_0), \\ u_{ijk}|y_{ijk} = 1 &\sim \text{Bernoulli}(\gamma_1), \end{aligned}$$

The sensitivity and specificity are usually regarded as properties of the diagnostic test. The values of these parameters are usually known for well established diagnostic tests.

2.2 The spatial piecewise exponential model

A Cox proportional hazards model (Cox, 1972) for the time-to-event t has hazard function

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}^T \tilde{\beta}),$$

where $\lambda_0(t)$ is the baseline hazard function, $\mathbf{x}$ is a vector of covariates, and $\tilde{\beta}$ is a vector of fixed effect coefficients. We consider a piecewise constant hazard function for analyses of the interval-censored disease surveillance data. The whole duration of study is partitioned into K intervals using the observation time $0 = \tau_0 < \tau_1 < ... < \tau_K < \infty$. Define the *k-th* interval as $(\tau_{k-1}, \tau_k]$ and assume that the hazard function is constant within each interval. To consider the spread of disease among the subjects within the same group, we propose a model where the hazard of a certain subject becoming diseased can be influenced by the status of the rest of the subjects within the same group and this influence is associated with the spatial distance between the subjects. Let λ_{ijk} be the hazard of subject j in group i becoming diseased in time interval $(\tau_{k-1}, \tau_k]$ given no event has happened by time τ_{k-1}. Then the conditional time-to-event $t_{ij}|y_{ij,k-1} = 0$ during time interval $(\tau_{k-1}, \tau_k]$ follows

an exponential distribution

$$t_{ij}|y_{ij,k-1}=0\sim\exp(\lambda_{\mathrm{ijk}}),\mathrm{t_{ij}}\in(\tau_{\mathrm{k-1}},\tau_{\mathrm{k}}].$$

Let $y_{ij',k-1}$ be the true disease status for subject $j^{'}\neq j$ at time τ_{k-1}, which has potential to affect the disease status of subject j at time τ_k. By using a log link function, the covariate effects can be incorporated into the hazard function, i.e., within each time interval $(\tau_{k-1},\tau_k]$,

$$(1)\quad \log\left(\frac{1}{\lambda_{\mathrm{ijk}}}\right)=\beta_0+\sum_{\mathrm{j'\neq j}}\left(\beta_1+\beta_2\exp(-\mathrm{d_{jj'}})\right)\mathrm{y_{ij',k-1}},$$

where β_0 is a regression parameter in the baseline hazard, β_1 is a regression parameter associated with $y_{ij',k-1}$ but not related to distance, and β_2 is a regression parameter associated with $y_{ij',k-1}$ and the distance between subject j and $j^{'}$. β_0 models the baseline negative log hazard of disease onset, i.e., effect of factors besides disease transmission among animals, for example, the air, the food, or the persons who work on the farm. β_1 guarantees that as the distance gets large, the effect of $y_{ij',k-1}$ to y_{ijk} does not decay to 0. Also, the exponential decay $\exp(-\mathrm{d_{jj'}})$ indicates that the influence of a diseased subject decreases as the distance between subjects gets further. We do not have a term of distance itself in model as we do not assume the hazard of disease onset would be affected by distance between pens only, regardless of their disease status. By a transformation on both sides of Eq.(1), the hazard function for time interval $(\tau_{k-1},\tau_k]$ can be expressed as:

$$\begin{aligned}\lambda_{ijk}&=\exp\left\{-\left(\beta_0+\sum_{j'\neq j}\left(\beta_1+\beta_2\exp(-\mathrm{d_{jj'}})\right)y_{ij',k-1}\right)\right\}\\(2)\quad&=e^{-\beta_0}\exp\left\{-\sum_{j'\neq j}\left(\beta_1+\beta_2\exp(-\mathrm{d_{jj'}})\right)y_{ij',k-1}\right\},\end{aligned}$$

where $e^{-\beta_0}$ can be interpreted as the baseline hazard and the covariate $y_{ij',k-1}$ is unknown and time-varying. In this way, the hazard function for subject j can be modeled using K parameters $\lambda_{ij1},...,\lambda_{ijK}$, each representing the risk of the sampled subject being diseased in one particular time interval. Because the risk is assumed to be piecewise constant and the distances among subjects are taken into consideration, the corresponding model is called a spatial piecewise exponential model. For each subject j such that $y_{ij0}=0$, the survival function for event time t_{ij} in interval $(\tau_{k-1},\tau_k]$ can be derived as:

$$S_{ijk}(t)\quad=\quad\exp\left\{-\left(\sum_{\mathrm{l<k}}\lambda_{\mathrm{ijl}}(\tau_{\mathrm{l}}-\tau_{\mathrm{l-1}})+\lambda_{\mathrm{ijk}}(\mathrm{t}-\tau_{\mathrm{k-1}})\right)\right\}$$

where λ_{ijk} is defined in Eq. (2). Then the probability density for the failure time of subject j in $(\tau_{k-1},\tau_k]$ can be derived as:

$$\begin{aligned}f_{ijk}(t)&=\frac{\partial}{\partial t}[1-S_{ijk}(t)]\\&=e^{-\sum_{l<k}\lambda_{ijl}(\tau_l-\tau_{l-1})}\lambda_{ijk}e^{-\lambda_{ijk}(t-\tau_{k-1})}.\end{aligned}$$

2.3 Likelihood of the latent disease process

Here we develop the likelihood of the latent disease process using the piecewise exponential model. Given that subject j is not diseased at time τ_{k-1}, $k\geq 1$, the t_{ij} follows an exponential distribution with λ_{ijk} defined in Eq. (2). Thus, the conditional density can be derived as

$$f(t_{ij}|y_{ij,k-1}=0)=\lambda_{ijk}e^{-\lambda_{ijk}(t_{ij}-\tau_{k-1})}$$

In particular, let $\mathbf{y}_{i.k}=(y_{i1k},...,y_{iJk})$ be a vector of the disease status of all subjects in group i at τ_k, the conditional probability of subject j being diseased at τ_k, $k\geq 1$ is:

$$\begin{aligned}&p(y_{ijk}=1|\mathbf{y}_{i.k-1};\tilde{\beta})\\=\;&I(y_{ij,k-1}=1)\\&+p(t_{ij}\in(\tau_{k-1},\tau_k]|y_{ij,k-1}=0)\cdot I(y_{ij,k-1}=0)\\=\;&I(y_{ij,k-1}=1)\\&+\int_{\tau_{k-1}}^{\tau_k}f(t_{ij}|y_{ij,k-1}=0)\mathrm{dt}\cdot\mathrm{I(y_{ij,k-1}=0)}\end{aligned}$$

Based on the assumption of conditional independence of $\mathbf{y}_{i.k}$ given $\mathbf{y}_{i.k-1}$, the likelihood for $\mathbf{y}_i=(\mathbf{y}_{i.0},\mathbf{y}_{i.1},...,\mathbf{y}_{i.K})$ can be derived as

$$\begin{aligned}f(\mathbf{y}_i;p_0,\tilde{\beta})&=f(\mathbf{y}_{i.K}|\mathbf{y}_{i.K-1};\tilde{\beta})f(\mathbf{y}_{i.K-1}|\mathbf{y}_{i.K-2};\tilde{\beta})\cdots\\&\quad f(\mathbf{y}_{i.1}|\mathbf{y}_{i.0};\tilde{\beta})f(\mathbf{y}_{i.0};p_0)\\(3)\qquad&=\prod_{j=1}^{J}[f(y_{ij0};p_0)\prod_{k=1}^{K}f(y_{ijk}|\mathbf{y}_{i.k-1};\tilde{\beta})].\end{aligned}$$

2.4 Joint likelihood function

Let $\mathbf{u}_i=(\mathbf{u}_{i.0},\mathbf{u}_{i.1},...,\mathbf{u}_{i.K})$ denote the observed outcome for all subjects in group i. For each group i, it is assumed that the observed outcomes $\mathbf{u}_i$ are conditional independent given latent true status $\mathbf{y}_i$ so that

$$\begin{aligned}f(\mathbf{u}_i|\mathbf{y}_i)&=\prod_{j=1}^{J}\prod_{k=1}^{K}f(u_{ijk}|y_{ijk})\\&=\prod_{j=1}^{J}\prod_{k=1}^{K}[(1-\gamma_0)^{u_{ijk}}\gamma_0^{1-u_{ijk}}]^{1-y_{ijk}}\\(4)\qquad&\quad\times[\gamma_1^{u_{ijk}}(1-\gamma_1)^{1-u_{ijk}}]^{y_{ijk}}.\end{aligned}$$

where the conditional distribution of u_{ijk} given y_{ijk} can be derived as from the misclassification model defined in (1). Let $\tilde{\mathbf{u}}=(\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_I)$ and $\tilde{\mathbf{y}}=(\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_I)$ denote the observed outcome and underlying true status for all subjects

in all groups, let p_0 and $\tilde{\beta} = (\beta_0, \beta_1, \beta_2)$ be the parameters of interest. Based on the assumption of independent groups, the joint likelihood function for the entire model is

$$\begin{aligned} L(\tilde{\mathbf{u}}, \tilde{\mathbf{y}}; p_0, \tilde{\beta}) &= \prod_{i=1}^{I} L_i(\mathbf{u}_i, \mathbf{y}_i; p_0, \tilde{\beta}) \\ &= \prod_{i=1}^{I} f(\mathbf{u}_i|\mathbf{y}_i) f(\mathbf{y}_i; p_0, \tilde{\beta}), \end{aligned} \tag{5}$$

where $f(\mathbf{y}_i; p_0, \tilde{\beta})$ and $f(\mathbf{u}_i|\mathbf{y}_i)$ are derived in Eq. (3), (4).

2.5 Bayesian estimation

The parameters of interest in this model are p_0 and $\tilde{\beta}$, where p_0 is the probability that an event happens at the beginning of sampling time and $\tilde{\beta} = c(\beta_0, \beta_1, \beta_2)$ are the regression parameters in the latent spatial piecewise exponential model. We propose to use a Bayesian method to estimate the model parameters. Because p_0 takes values in $[0, 1]$, a conjugate Beta prior distribution $f(p_0)$ is used for p_0. The Normal distribution priors $f(\beta_i)$ are used for the regression parameters $\beta_i, i = 0, 1, 2$. Based on the joint likelihood function derived in (10), the joint posterior distributions for p_0, β_0, β_1 and β_2 can be derived as

$$\begin{aligned} f(p_0, \tilde{\beta}, \tilde{\mathbf{y}}|\tilde{\mathbf{u}}) &= \prod_{i=1}^{I} f(\mathbf{u}_i, \mathbf{y}_i|p_0, \beta_0, \beta_1, \beta_2) \\ &\quad \times f(p_0) f(\beta_0) f(\beta_1) f(\beta_2) \end{aligned}$$

Non-informative priors are used for all model parameters in this paper. Informative priors can be applied if prior knowledge is available for one or more parameters. Markov chain Monte Carlo (MCMC) is used to generate a sequence of draws from the posterior distribution of the parameters. The unknown true disease status $\tilde{\mathbf{y}}$ is also simulated in the MCMC algorithm along with the parameters. Inference for the model parameters is then based on the draws from the posterior distribution. The MCMC method is done using freely-distributed software JAGS version 4.0 (Plummer, 2015).

3. THE ANALYSIS OF PRRSV SURVEILLANCE DATA

In this section, the proposed model is applied to a dataset based on oral fluid samples collected from three barns on one swine farm. Surveillance data facilitates the efficient use of resources for the control of infectious disease and is essential for control/elimination programs. Oral fluid sampling and testing using nucleic acid- or antibody-based assays is one approach used for increasing the efficiency and reducing the cost of surveillance in swine herds (Ramirez et al., 2012). Oral fluid samples are easily collected from pens of pigs by allowing them to chew on cotton rope suspended in the pen for 20-30 min, manually extracting the fluid from the rope, and decanting the sample into a tube for submission to the laboratory. In this study, samples collected from each pen were tested for PRRSV nucleic acids using polymerase chain reaction (PCR) -based assay. In this situation, the pen is treated as the subject unit with a fixed location in a 2×18 matrix. The event of interest is defined as a PCR-positive, i.e., the assay detects PRRSV nucleic acids from one or more of the individual pigs in the pen.

The oral fluid samples were collected in a weekly schedule for a consecutive of 8 weeks (week 0 to week 8, 9 time points) from 3 wean-to-finish barns on one finishing site. There are 36 pens within each barn and around 25 pigs in each pen. The samples were collected in a way as described above, with the sampling unit being the pen. This provided a total of $3 \times 36 \times 9 = 972$ samples. The samples were completely randomized and then tested for PRRSV RNA. Let u_{ijk} denote the diagnostic test outcomes for pen j in barn i at sampling time τ_k, $u_{ijk} = 1$ if the test outcome is positive and 0 if negative. The test outcomes are recorded with misclassification because of the imperfect sensitivity and specificity of the diagnostic test. Sensitivity measures the proportion of actually diseased sampling units which are correctly identified as diseased, whereas specificity measures the proportion of actually healthy sampling units which are correctly identified as healthy. So the diagnostic sensitivity and specificity of the assay refer to γ_1 and γ_0, respectively. Reference values for γ_1 and γ_0 were 0.9 and 0.98 from related studies (Olsen et al., 2013). The number of diseased pens in each barn (among 36 pens) at each sampling week are shown in Table 1. An individual response profile (IRP) plot in Figure 2 shows the change of test outcomes for each pen overtime.

From Table 1, it can be seen that the initial probability for a pen being PRRSV-positive was very low, with an average about $5/108 \approx 0.046$. As the virus spread over time, more pens became infected, i.e., tested PRRSV PCR-positive. That is, few pens were infected initially, but the virus continually spread to surrounding pens until, gradually, all pens were virus-positive at the end of the study period. Figure 3 illustrates the pattern of spread among pens by barn. Based on the test outcomes, the mean time to disease for the three barns were 3.81 weeks (A), 4.75 weeks (B) and 6.42 weeks (C). The proposed model was applied to study the time to disease t_{ij} for each pen within the building, with $p_0 = p(y_{ij0} = 0)$ and the hazard $\lambda_k = e^{-\beta_0} \exp\{-\sum_{\mathrm{j}' \neq \mathrm{j}} (\beta_1 + \beta_2 \exp(-\mathrm{d}_{\mathrm{jj}'})) \mathrm{y}_{\mathrm{ij}',\mathrm{k}-1}\}$ for $t_{ij} \in (\tau_{k-1}, \tau_k]$.

A Bayesian approach is used to estimate the model parameters, similar to the simulation studies. Estimation was based on 30,000 MCMC iterations after disregarding the first 20,000 MCMC draws as burn-in. The posterior means, standard deviations and 95% Bayesian credible intervals (CI) are reported in Table 2. The 95% Bayesian credible intervals do not include 0, indicating that the parameter estimates are statistically significant. The constructed model

Table 1. Number of diseased pens in each Barn at each Time point

Week	Barn	♯Diseased.Pens	Barn	♯Diseased.Pens	Barn	♯Diseased.Pens
0	A	3 (36)	B	2 (36)	C	0 (36)
1	A	6 (36)	B	2 (36)	C	1 (36)
2	A	7 (36)	B	2 (36)	C	1 (36)
3	A	15 (36)	B	5 (36)	C	1 (36)
4	A	16 (36)	B	13 (36)	C	0 (36)
5	A	30 (36)	B	24 (36)	C	3 (36)
6	A	36 (36)	B	36 (36)	C	8 (36)
7	A	36 (36)	B	36 (36)	C	34 (36)
8	A	36 (36)	B	36 (36)	C	36 (36)

has successfully provided evidence for the spread of disease and the spatial relationship among the pens. In particular, the initial probability of a pen being infected is estimated to be $\hat{p}_0 = 0.048$. The baseline hazard of a pen being diseased within a time interval is estimated as $e^{-\hat{\beta}_0} = 0.014$. Both parameters β_1 and β_2 associated with the covariates are negative, which indicates an increase in the hazard of a pen being infected when other pens in the same barn are infected. It is estimated that the existence of an infected pen in a barn will increase the hazard of healthy pens in the same barn to become diseased by a multiplicative factor of $\exp(0.219 + 3.512\ \exp(-\mathrm{d}_{\mathrm{jj}'}))$. It is clear to see that the influence decreases as the distance between two pens becomes larger.

Table 2. Posterior results for parameters in the analysis of PRRSV data

	Posterior mean	Posterior S.D.	95% CI
p_0	0.048	0.0208	[0.0157, 0.0958]
β_0	4.236	0.3525	[3.6290, 4.8795]
β_1	-0.219	0.0361	[-0.2885, -0.1583]
β_2	-3.512	0.4262	[-4.2467, -2.6676]

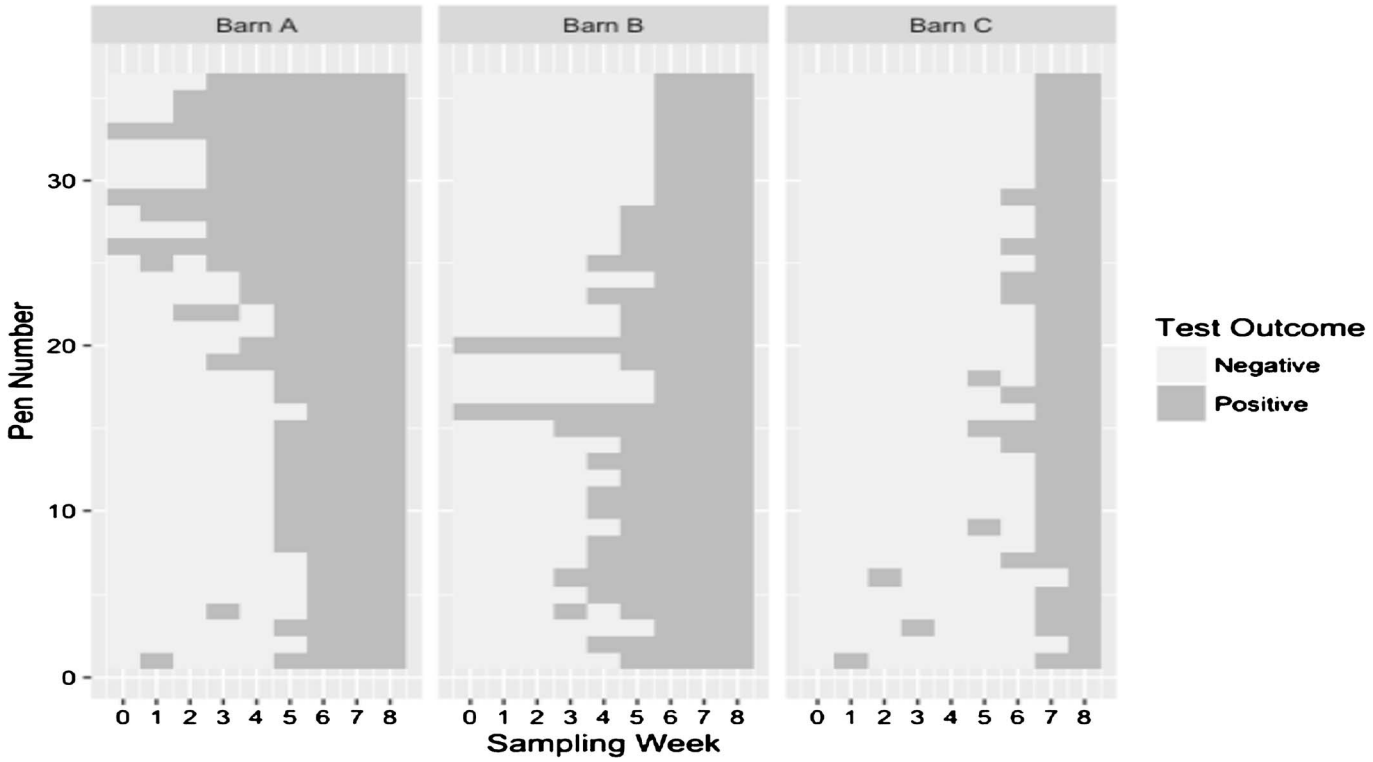

Figure 2. Individual response plot (IRP) for diagnostic test outcomes of each pen overtime for all three barns based on the PRRSV surveillance data.

4. SIMULATION STUDY

A simulation study was performed to investigate the performance of the proposed model. To mimic the true data structure in the PRRSV surveillance study, we set the group size $I = 3$ and number of subjects $J = 36$ within each group. All groups have equal size and structure, the subjects within the group have fixed locations and are arranged in a 2×18 matrix. The distance between the subjects can be calculated as the Euclidean distance. Time-to-event data are simulated for each subject following the proposed spatial piecewise exponential model. The observed outcomes are then simulated with the misclassification parameters. The proposed Bayesian approach is then applied to estimate the model parameters. One thousand simulations were run for the parameter configuration.

The initial probability of an event occurrence is set to $p_0 = 0.05$. Within each time interval $(\tau_{k-1}, \tau_k], k \geq 1$, the hazard is generated as in (3), where model parameters are set to be $\beta_0 = 4, \beta_1 = -0.2, \beta_2 = -3$. According to the model specification, the time-to-event data were generated as follow. The sampling period is divided into 9 intervals $\tau_0 < \tau_1 < ... < \tau_9$, where $\tau_k - \tau_{k-1} = 1$ for $k = 1, ..., 9$. For each subject j in group i, first generate true status y_{ij0} from a Bernoulli(p_0) distribution at the initial sampling point τ_0. If $y_{ij0} = 1$, the event happens to the subject at the beginning of the study and thus $y_{ijk} = 1$ for $k \geq 1$. Otherwise simulate t_1 from an exponential distribution with rate λ_{ij1}, where λ_{ij1} is calculated from (3). If $t_1 \leq 1$, which means the event happened to the subject before τ_1 and after τ_0, let $t_{ij} = t_1$, thus $y_{ij0} = 0$ and $y_{ijk} = 1$ for $k \geq 1$. If $t_1 > 1$, then $y_{ij0} = y_{ij1} = 0$ and we generate t_2 from an exponential distribution with rate λ_{ij2}, where λ_{ij2} is calculated from (3). Similarly, if $t_2 \leq 1$, let $t_{ij} = 1 + t_2$ and the event happened to the subject before sampling time τ_2, thus $y_{ij0} = y_{ij1} = 0$ and $y_{ijk} = 1$ for $k \geq 2$. If $t_2 > 1$, $y_{ijk} = 0$ for $k = 0, 1, 2$ and t_3 is generated from an exponential distribution with rate λ_{ij3}. Similarly in this way, the failure time t_{ij} for each subject can be generated. With the predetermined misclassification parameters $\gamma_1 = 0.9, \gamma_0 = 0.98$, the observed outcome u_{ijk} then can be generated from the following Bernoulli distributions:

$$\begin{aligned} u_{ijk}|(y_{ijk} = 1) &\sim \text{Bernoulli}(\gamma_1), \\ u_{ijk}|(y_{ijk} = 0) &\sim \text{Bernoulli}(1 - \gamma_0), \end{aligned}$$

where $i = 1, 2, 3, j = 1, 2, ..., 36, k = 0, 1, 2, ..., 9$.

One thousand simulated sets of data were generated and analyzed by the proposed Bayesian approach for parameter estimation. The prior for p_0 was chosen as Beta$(0.5, 0.5)$, which is a non-informative conjugate prior for p_0 on interval $(0, 1)$. The priors for the regression parameters β_0, β_1 and β_2 were chosen as non-informative $N(0, 1000^2)$. For each simulated data set, model implementation is based on 30,000 MCMC iterations after disregarding the first 20,000 MCMC realizations as the burn-in procedure for the algorithm. Three chains are run to check for convergence of the MCMC algorithm based on the scale reduction factors. Based on 1000 simulation trials, the bias, standard errors and root mean square errors (RMSE) of posterior means as parameter estimators are calculated and reported in Table 3 (Model 1).

To investigate the effect of misclassification, a model without misclassification is also considered for comparison by treating the observed status as true status, i.e. $u_{ijk} = y_{ijk}$ for all i, j, k. In this case, the exact time interval where the disease occurred is assumed to be known and no latent class for disease is needed. Parameters in this simplified model can be estimated using either Bayesian method or maximum likelihood method. For Bayesian method, priors for model parameters are chosen the same as in Model 1. Based on the same 1000 simulated datasets, the parameter estimation bias, standard deviation and RMSE are also reported in Table 3 (Model 2).

The results in Table 3 suggest that the posterior means for the initial probability p_0 and regression parameters

Table 3. Bias, standard deviation and RMSE for parameter estimations in the model with misclassification (Model 1) and model without misclassification (Model 2) based on 1000 simulations

Model	Parameter	True Value	Bias	S.D.	RMSE
Model 1	p_0	0.05	0.004	0.0225	0.0228
	β_0	4	0.03	0.3619	0.3633
	β_1	-0.2	-0.004	0.0368	0.0370
	β_2	-3	0.03	0.5841	0.5845
Model 2 (Bayesian)	p_0	0.05	0.017	0.0232	0.0289
	β_0	4	-0.538	0.3499	0.6413
	β_1	-0.2	0.048	0.0341	0.0589
	β_2	-3	0.759	0.4265	0.8707
Model 2 (Likelihood)	p_0	0.05	0.013	0.0236	0.0269
	β_0	4	-0.637	0.3328	0.7183
	β_1	-0.2	0.043	0.0399	0.0587
	β_2	-3	0.736	0.4649	0.8704

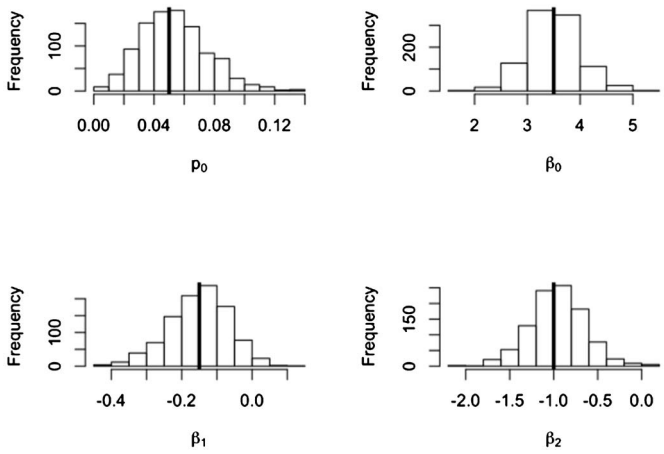

Figure 3. Histograms of posterior means for model parameters obtained from 1000 simulations. The bold line indicates the true value used for generating the data.

$\beta_0, \beta_1, \beta_2$ are approximately unbiased estimators under the proposed model. The histograms of the posterior means for the parameters appear unimodal, bell shaped and centered around the true parameter values (Figure 3). However, if misclassification is not considered, the estimation results can be very biased as shown in Table 3 (Model 2).

5. CONCLUSIONS AND FUTURE WORK

We have proposed a latent spatial piecewise exponential model to study interval-censored disease surveillance data with time-varying covariates and misclassification. This proposed model allows researchers to model disease spread among units and relate the transmission rate to the distance between units, even when the true disease status is unobserved. The model is interpretable, flexible and straightforward to implement. Our model allows the assessment of time-varying covariates, which were the latent disease status in our application. Simulation studies also show that when observed outcomes are with misclassification, the proposed model works better in parameter estimation than the model without misclassification. When applied to the PRRSV surveillance data, the model results in significant model parameter estimates, thereby providing strong evidence of distance related disease spread among pens within a barn. We use exponential distance decay due to its popularity in modeling of spatial effects. Other functional forms of distance decay is possible. We have also tried the reciprocal decay; the results from simulation studies and real data analyses remains consistent. Once the disease spread pattern is captured using the proposed model, it can be utilized for developing sampling guidelines, e.g., optimal sample size, sampling frequency and sample allocation to maximize the power of disease detection at a minimal cost.

Some extensions of the model can be considered in future work. In the application of PRRSV surveillance data, samples were taken weekly so the time intervals are equally spaced as $\tau_k - \tau_{k-1} = 1$. This model however is not limited to equally spaced time intervals, and can be generalized easily to situations that different subjects are sampled at different regular or irregular time points. Another extension can be the consideration of recurrent event data. In this paper, it is assumed that the event happens only once per unit, that is, when the unit is diseased, the status stays the same until the end of the study. This assumption is appropriate for studying of the data of PRRSV. Another consideration is diseases in which recovery does not preclude reinfection, such as influenza. In such instances, analyses for recurrent event data need to be developed. A general idea is to develop models for a sequence of events, one for each disease occurrence and/or recovery process. Furthermore, whereas there are only three groups and the groups are assumed to be independent from each other in this PRRSV surveillance data, modern production systems have more hierarchies of structures, e.g., company → farm → barn → pen. These could be taken into account when evaluating disease spread at levels higher than the pen by collecting data at multiple levels. Random effects can also be added to the model to capture random variation at higher levels. All the above topics are the subjects of the on-going research.

ACKNOWLEDGMENTS

This study was funded in part by Checkoff Dollars administered through the National Pork Board (Grants #13-157 and #17-174), Des Moines, Iowa (USA).

Received 25 May 2017

REFERENCES

Berry, S.M., Berry, D.A., Natarajan, K., Lin, C.S., Hennekens, C.H., Belder, R. (2004) Bayesian survival analysis with nonproportional hazards: Meta-analysis of pravastatin-aspirin. *Journal of the American Statistical Association* **99**: 36–44. MR2061886

Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**(4): 845–854. MR0872963

Friedman, M. (1982) Piecewise exponential models for survival data with covariates. *The Annals of Statistics* **10**(1): 101–113. MR0642722

García-Zattera, M.J., Hara, A. and Komárek, A. (2016). A Flexible AFT Model for Misclassified Clustered Interval-Censored Data. *Biometrics* **72**: 473–483. MR3515774

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin. D.B. (2004) *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC. MR2027492

Goodman, M.S., Lib, Y., Tiwari, R.C. (2011) Detecting multiple change points in piecewise constant hazard functions. *Journal of Applied Statistics* **38**: 2523–2532. MR2845706

Holtkamp, D.J., Kliebenstein, J.B., Neumann, E.J., Zimmerman, J.J., Rotto, H.F., Yoder, T.K., Wang, C., Yeske, P.E., Mowrer, C.L., Haley, C.A. (2013). Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States pork producers. *Journal of Swine Health and Production* **21**(2): 72–84.

Huang, J., Wellner, A.J. (1997). Interval censored survival data: a review of recent progress. In Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis. Lect. Notes Stat. **123**: 123–169.

Lin, H., Wang, C., Liu, P., Holtkamp, D.J. (2013). Construction of disease risk scoring systems using logistic group lasso: application to porcine reproductive and respiratory syndrome survey data. *Journal of Applied Statistics* **40**(4): 736–746. MR3047315

Lindsey, J.C. and Ryan, L.M. (1998). Tutorial in biostatistics methods for interval-censored data. *Statistics in Medicine* **17**(2): 219–238.

Lyles, R.H., Tang, L., Superak, H.M., King, C., Celentano, D., Lo, Y. and Sobel, J.D. (2011). Validation Data-Based Adjustments for Outcome Misclassification in Logistic Regression: An Illustration. *Epidemiology* **22**: 589–597.

Mckeown, K. and Jewell, N.P. (2010). Misclassification of current status data. *Lifetime Data Analysis* **16**: 215–230. MR2608286

Olsen, C., Wang, C., Hennings, J.C., Doolittle, K., Harmon, K., Abate, S., Kittawornrat, A., Lizano, S., Main, R., Nelson, E., Otterson, T., Panyasing, Y., Rademacher, C., Rauh, R., Shah, R., Zimmerman, J. (2013). Probability of detecting Porcine reproductive and respiratory syndrome virus infection using pen-based swine oral fluid specimens as a function of within-pen prevalence. *Journal of Veterinary Diagnostic Investigation* **25**(3): 328–335.

Plummer, M. rjags: Bayesian graphical models using MCMC. R package version 4-0. (2015). Available at: http://cran.r-project.org/package=rjags.

Ramirez, A., Wang, C., Prickett J.R., Pogranichniy, R., Yoon, K.J., Main, R., Johnson, J.K., Rademacher, C., Hoogland, M., Hoffman, P., Kurtz, A., Kurtz, E., Zimmerman, J. (2012). Efficient surveillance of pig populations using oral fluids. *Preventive Veterinary Medicine* **104**: 292–300.

Yi, G.Y., Ma, Y., Spiegelman, D., Carroll, R., J. (2015). Functional and Structural Methods With Mixed Measurement Error and Misclassification in Covariates. *Journal of the American Statistical Association* **110**(510): 681–696. MR3367257

Zimmerman, J.J., Benfield, D.A., Dee, S.A., Murtaugh, M.P., Stadejek, T., Stevenson, G.W., Torremorell, M. (2012). Porcine reproductive and respiratory syndrome virus (porcine arterivirus). In: Zimmerman, J. J., Karriker, L. A., Ramirez, A., Schwartz, K. J., Stevenson, G. W. *Diseases of Swine, 10th Edition.* Wiley-Blackwell, Hoboken NJ, pp. 461–486.

Yaxuan Sun
3414 Snedecor Hall
Iowa State University
USA
E-mail address: evansun@iastate.edu

Chong Wang
3414 Snedecor Hall
Iowa State University
USA
E-mail address: chwang@iastate.edu

William Q. Meeker
2109 Snedecor Hall
Iowa State University
USA
E-mail address: wqmeeker@iastate.edu

Max Morris
1121H Snedecor Hall
Iowa State University
USA
E-mail address: mmorris@iastate.edu

Marisa L. Rotolo
VMRI 01
Iowa State University
USA
E-mail address: mrotolo@iastate.edu

Jeffery Zimmerman
VMRI 01
Iowa State University
USA
E-mail address: jjzimm@iastate.edu