Environmental and Ecological Statistics (2019) 26:17–45
https://doi.org/10.1007/s10651-019-00418-3

Bayesian-based survival analysis: inferring time to death in host-pathogen interactions

Sama Shrestha[1] · **Bret D. Elderd**[2] · **Vanja Dukic**[1]

Received: 4 May 2017 / Revised: 12 January 2019 / Published online: 8 February 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

The standard approach to modeling survival times, or more generally, time to event data, is often based on parametric assumptions that may not fit the data collected well. One of the goals of this article is to discuss and compare several commonly used parametric and non-parametric, as well as a Bayesian semi-parametric method for survival data. We do so in the context of the data from an experimental system where insect herbivores become infected when consuming a lethal virus along with the plant on which the virus resides. We used data collected on individual insects that were fed known doses of virus along with varying genotypes of a single plant species (soybean), to compare how the insect's diet affects its time to death. Through hazard characterization and model selection, we found that the flexible semi-parametric analysis is better at describing the time-to-death data while maintaining a relatively parsimonious form. Unlike the standard parametric and non-parametric approaches, the Bayesian semi-parametric approach better captured the rapid decline in the hazard function after a window of time where the host was most vulnerable to the virus. For our study system, being able to accurately model time to death and quantify how plant genetics affects within-insect disease processes allows us to gain a better understanding of the host-pathogen interaction at an individual level. While we show the appropriateness of the Bayesian semi-parametric approach for infection data, this method readily applies to data sets concerned with characterizing a time until any event.

Keywords Baculovirus · Bayesian semi-parametric analysis · Fall armyworm · Survival analysis · Time to death · Within-host

Handling Editor: Bryan F. J. Manly.

B. D. Elderd, V. Dukic: Joint senior authors.

Extended author information available on the last page of the article

1 Introduction

Survival analysis has a long history both within and outside ecology (Cox and Oakes 1984; Muenchow 1986). Generally, survival analysis is used to analyze time to event data such as "failure" or "death" (Kleinbaum and Klein 2006). However, in many data sets, not all event times are observed. Thus, survival data often consists of cases in which the event was not observed up to a specific point in time (either the individual was removed from the study or the study ended before the event occurred). Such data are known as censored data. While standard statistical techniques (e.g., those for regression modeling and density estimation), can sometimes be useful for such data, the presence of censoring makes them generally inapplicable. In this paper, we present a relatively recent flexible methodological approach for survival analysis, Bayesian Multiresolution Hazard (MRH), and compare it to several commonly used parametric and non-parametric approaches. We further investigate its ability to assess whether differences in plant food quality affect the time to insect death in a tritrophic interaction between a pest herbivore, its lethal pathogen, and the plant that the herbivore consumes.

The effect of external biotic factors such as resource quality on host-pathogen interactions remains an open-ended and well-recognized question in disease ecology (Lively et al 2014). An important facet of how the tritrophic interaction between the resource, a herbivore and its pathogen affect long-term disease dynamics includes how long it takes for exposed individuals to become infectious and pass along the virus to a new round of susceptible individuals. For lethal viruses where the next round of viral particles are only spread upon a host's death, the point of interest is the time at which the host dies. Thus, we seek to model host mortality time and the effects that factors such as diet can have on it.

A number of researchers have documented how infection rates of larval Lepidopterans due to baculoviruses are influenced by the resource quality of the host plant (Richter et al 1987; Keating and Yendol 1987; Foster et al 1992; Farrar and Ridgway 2000; Raymond et al 2002; Ibrahim Ali et al 2002; Szewczyk et al 2006; Elderd et al 2013). These interactions are often driven by the degree to which the host plant defends itself via plant secondary metabolites against herbivory. These plant secondary metabolites affect host plant quality, which can indirectly affect the host's response to being challenged by a pathogen. However, these studies often ignore within-host processes and focus on population-level dynamics, or simply whether or not an individual becomes infected. Given that the time between infection and death can have important implications for population-level processes (Kennedy et al 2014), understanding how plant defenses affect the host mortality time is a key question.

To describe the within-host infection process using host mortality time and the influences of external biotic processes, we focused on analyzing a series of experiments where individual Lepidopteran larvae were infected with a baculovirus. We were interested in investigating the effects of explanatory variables such as host resources and how they affect the time to death after ingestion of a lethal pathogen. We analyzed these data using several common survival analysis approaches, including the MRH, to highlight the potential benefits as well as shortfalls of these methods. Specifically, we used parametric, non-parametric, and Bayesian semi-parametric frameworks. By

looking at a wide variety of survival models, we obtained a good characterization for the hazard and survival function of our host study organism, the fall armyworm *Spodoptera frugiperda*. We found that the Bayesian semi-parametric MRH models performed better at capturing the characteristics of the distribution associated with the time to death. This reflects the greater flexibility of the MRH approach in comparison to more commonly used parametric models.

2 Methods

2.1 The system

The fall armyworm is a multivoltine (i.e., multiple generations within a single year) polyphagous migratory species with non-overlapping generations. The life cycle of the insect begins when adult females lay eggs in clusters on a variety of substrates. After the eggs hatch, there are six larval instars (developmental stages) requiring 14–30 days to reach pupation (Pitre and Hogg 1983). Adults emerge to mate and continue the fall armyworm life cycle. Outbreaks of the fall armyworm, which have been recorded as early as 1845 (Hinds and Dew 1915), can be quite large and wide-spread (Pair et al 1991).

Baculoviruses are ubiquitous in nature and infect a wide-range of insect species (Miller 1997) including the fall armyworm. Baculovirus infections begin when foliage contaminated with baculovirus occlusion bodies (OBs) are consumed by a larva (Cory and Myers 2003). The OBs contain multiple virions surrounded by a protein coat, which dissolves in the host mid-gut. If enough OBs are consumed, a fatal infection occurs. The virus then replicates within the host producing millions of viral particles until the baculovirus triggers host liquefaction (Cory and Myers 2003). Transmission occurs when OBs released upon liquefaction contaminate the foliage on which susceptible hosts are feeding, and the infection cycle continues (Dwyer et al 2000). For the fall armyworm, the species-specific baculovirus, *Spodoptera frugiperda* multicapsid nucleopolyhedrovirus (SfMNPV), can infect up to 50–60% of the individuals in fall armyworm infested areas (Fuxa 1982). Thus, SfMNPV represents an important source of mortality in this system (Richter et al 1987).

As agricultural pests, fall armyworms readily feed on a number of different crops including soybeans (Richter et al 1987; Sparks 1979) with later instars causing the majority of crop damage (Sparks 1979). Soybean genotypes or isolines vary in the amounts of plant secondary metabolites that they produce (Underwood et al 2002; Bi and Felton 1995) including proteinase inhibitors and oxidative enzymes (Bi and Felton 1995; Botella et al 1996). These compounds affect insects directly, by altering feeding behavior, damaging midgut tissues, and interfering with digestive processes. These chemical compounds also affect insects indirectly by interacting with baculoviruses in the insect midgut (Hoover et al 1998). Soybeans produce these chemical defenses either constitutively or via induction due to herbivores feeding on the plant. Given the wide range in both the way in which the chemical defenses are expressed and the amount of chemical defenses produced by individual isolines, soybean serves as an ideal plant to examine how resource quality affects time to death. Additionally,

soybeans self pollinate and, thus, produce genetically similar offspring. By using soybeans, we were able to examine the time to death once infected without being concerned about differences in plant quality due to genetic variation.

2.2 Experimental design

We carried out a series of laboratory experiments in order to determine how differences in leaf tissue/resource quality affect the time to death of lethally infected fall armyworms. Newly hatched larvae were transferred to two oz. plastic cups, containing approximately one half oz. of a wheat-based artificial diet. The larvae were reared on the artificial diet at 25 °C until they reached the third instars with head capsules that had begun to slip forward (a sign that fourth instar is imminent). The larvae were then removed from their diet and starved overnight. This ensured that all larvae were at the same development stage since larval age affects the susceptibility of the larvae to the virus (Hoover et al 1998). We chose to examine time to death using fourth instars since the fourth instars not only cause a great deal of defoliation but also play an important part in the disease transmission process (Elkinton and Liebhold 1990; Elderd and Reilly 2014).

We chose nine different genotypes of soybeans along with a diet control to examine how resource quality affects time to death. Three of the genotypes could be considered non-inducible such that they do not upregulate plant secondary metabolites when eaten by herbivores (Underwood et al 2000, 2002). The remaining six genotypes have been shown to upregulate plant secondary metabolites due to herbivore damage and can be considered inducible (Underwood et al 2000, 2002). The soybeans were grown at 28.9 °C with a 16 h day, 8 h night cycle until their leaves were at the two trifoliate stage.

For the experiment, we cut out 1 cm^2 of undamaged leaf tissue from multiple trifoliate leaves for each genotype and put a standard artificial diet cube on the leaf disks. These were placed in an empty two oz. cup. A 3 µl droplet of water with 10^5 OBs was then placed on the diet cube. 10^5 represents a rough estimate of LD95, or the lethal dose at which 95% of those infected die when fed the virus on a cube of artificial diet (Elderd, unpublished data). Controls consisting of 3 µl of water without OBs were used to check for background contamination in the lab, to assure that no mode of transmission other than that due to diet was present. The recently molted fourth instar larvae were then placed in the cup and allowed to feed for 24 h. We also had a diet group where the larvae were fed a virus inoculated diet cube without a leaf disk. Larvae that did not consume the whole leaf disks along with the cube of infected diet were discarded and not included in the analysis. All individuals included in the analysis were placed back on artificial diet and reared in an environmental chamber at 28.9 °C under a 16 h day and 8 h night. We counted the number of dead larvae every 12 h, which was confirmed either visually or under a light microscope where OBs are clearly visible (Cory and Myers 2003). The total sample size from our experiment was 555 individual larvae, with 433 larvae dying from infection (cause of death verified) during the experiment, and 122 surviving at least until the end of the experiment.

3 Modeling fall armyworm time to death

The experiment described above gives rise to data of the form

$$(t_{ij}, c_{ij}); \quad i = 1, \ldots, k; \quad j = 1, \ldots, 10,$$

where t_{ij} represents the time of death or time of censoring (within 12-h intervals) of the ith larva in the jth soybean genotypic group (with the artificial diet being treated as the 10th group). We have right censored mortality data for 122 larvae that survived the experiment. The censoring indicator, c_{ij} is a binary variable with 0 denoting right censoring (death not observed within the experimental time frame) and 1 otherwise (death observed within the experimental time frame).

In our study, the random variable of interest is the continuous mortality time (T) of each of the larvae within each plant genotype. In addition to describing the distribution of T, we also aim to look at variability in time to death with respect to the plant genotype on which the larvae consumed the virus. There are several ways of analyzing survival time data, the overall goal being to summarize the main features of the distribution and examine the effects of explanatory variables (Dobson 2002). For our data, we will estimate the survival and hazard functions of T using parametric, non-parametric models and semi-parametric methods. Parametric models require the specification of a known family of probability distributions for the time to death while non-parametric models do not assume a specific probability distribution. Semi-parametric models contain both a parametric form and a non-parametric form in its model definition.

3.1 Non-parametric analysis

The simplest and most commonly used non-parametric method for visualizing censored survival data is the Kaplan–Meier curve (Kaplan and Meier 1958). The Kaplan–Meier estimator uses both censored and non-censored information in the sample based on the following formula:

$$\hat{S}(t) = \prod_{h:\, t_h \le t} \frac{n_h - d_h}{n_h}, \tag{1}$$

where t_h is time at which at least one death happened, n_h is the number of individuals alive just before time t_h and d_h is the number of deaths at time t_h.

3.1.1 Kaplan–Meier estimates for the fall armyworm data

We estimated Kaplan–Meier survival curves and their point-wise confidence bands estimated using Greenwood's formula (Greenwood 1926) for each of the soybean genotypes (Fig. 1, Table 1). Figure 1 shows that the survival probability in each group decreases steadily after a few days into the infection process. Genotypes Davis, Gasoy and Stonewall have a few additional deaths after a long time interval towards the end of the experiment. However, there is a notable difference in the diet group which represents the non-soybean diet consumption of the virus: half the population survives

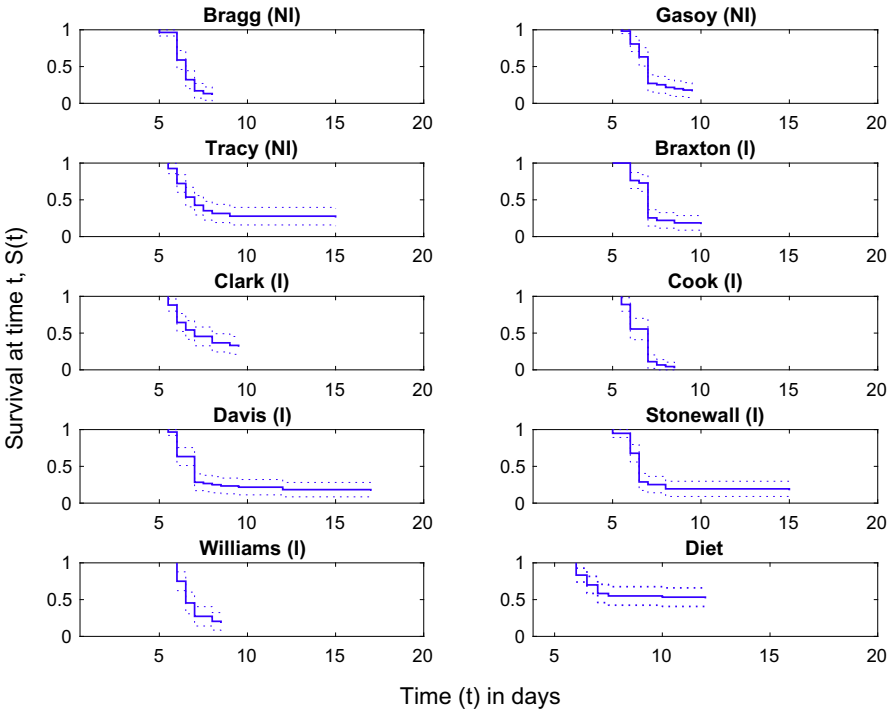


Fig. 1 Kaplan–Meier survival curves and their 95% confidence bands for the fall armyworm mortality data. 'NI' stands for non-inducible genotypes and 'I' for inducible genotypes (Underwood et al 2000). The diet group fed on virus-laced artificial diet only and did not ingest any soybean leaves

at the end as seen in the figure as the survival curve levels off at the 0.5 value for $S(t)$. Based on this exploration of the data, there seems to be a need to model the diet group separately from other groups. This also gives rise to the hypothesis that the presence of soybean genotype in the diet of the larvae significantly affects its mortality rate. Additionally, it seems that the genotypes themselves differ in terms of fall armyworm mortality.

3.2 Parametric analysis

While the Kaplan–Meier estimator works well as an exploratory tool, survival models are better capable of assessing systematic differences between groups of covariates. Furthermore, they provide greater flexibility with regard to model comparison and selection. They also enable the inclusion of covariate effects in our model such as genotypes. Survival models are also able to precisely account for different types of censoring. Specifically, in our dataset the exact times of deaths are unknown; the only information we have is that they happened between data collection points, within 12 h intervals. This is known as interval censoring (Kleinbaum and Klein 2006). Survival models will also be able to account for the extra uncertainty due to the imprecise time observations.

Table 1 Table of estimated Kaplan–Meier mean and median survival times (in days) for each genotype along with the respective sample sizes (Brookmeyer 2014)

Genotype	Median	Mean	Sample size
Williams	6.42	7.15	44
Stonewall	6.23	7.42	59
Gasoy	6.68	7.26	57
Bragg	6.17	7.01	56
Braxton	6.74	7.20	60
Clark	6.74	7.37	59
Davis	6.38	7.38	60
Tracy	6.67	7.41	54
Cook	6.13	6.62	45
Diet	–	9.49	60

The median survival time for the Diet group is not calculated because the diet survival function does not fall below 0.5 in our data set

To start, we consider distributions drawn from the extensive family of generalized gamma distributions such as Weibull, lognormal and gamma which are some of the most commonly used distributions for parametric modeling of time-to-event data (Cox et al 2007). Additionally, we look at various ways of building these parametric models ranging from using individual parametric distributions for each genotype to including random effects or group effects in our models.

3.2.1 Fixed effects model

We begin by considering the simplest model where each soybean genotype mortality dataset is fitted individually by a common probability distribution, with no relationship between the different genotypes. Each group $j \quad (j = 1, \ldots, 10)$, has its own set of parameters (Θ_j) that parametrize its time to death distribution. The likelihood function for this "fixed effects" model can be expressed as:

$$\prod_{j=1}^{10} \prod_{i=1}^{n_j} L\left(\Theta_j | t_{ij}, c_{ij}\right) \propto \prod_{j=1}^{10} \prod_{i=1}^{n_j} f\left(t_{ij} \mid \Theta_j\right)^{c_{ij}} S\left(t_{ij} \mid \Theta_j\right)^{1-c_{ij}}, \tag{2}$$

where n_j is the number of worms in group j, $f(t_{ij} \mid \Theta_j)$ is the probability density function of time to death (only playing a role in the likelihood when the observation is uncensored and $c_{ij} = 1$), and $S(t_{ij} \mid \Theta_j)$ is the corresponding survival function (only playing a role in the likelihood when the observation is censored and $c_{ij} = 0$), both depending on the group-specific set of parameters Θ_j.

In our study the times t_{ij} are not observed continuously, but rather are reported only in half-day (12-h) increments, where $t_{ij} = z$ indicates that the event happened sometime during the previous 12 h $(z - 0.5, z]$ in case of non-censored events. For censored events, $t_{ij} = z$ would simply indicate that the censoring occurred at that time. Thus, we cannot use the above formula (2) directly. Instead, we can explicitly

model the interval censoring in the likelihood as follows:

$$\prod_{j=1}^{10}\prod_{i=1}^{n_j}\left(\int_{t_{ij}-0.5}^{t_{ij}} f(s \mid \Theta_j)\, ds\right)^{c_{ij}} S(t_{ij} \mid \Theta_j)^{1-c_{ij}}. \tag{3}$$

We implement three different choices of the generalized gamma density functions $f(t_{ij} \mid \Theta_j)$, namely Weibull, lognormal and gamma, below.

Weibull Distribution Fixed Effects Model We first consider the Weibull distribution, which can capture accelerated failure time with their hazard functions either being monotonically increasing or decreasing based on the value of scale and shape parameters. Thus, it could be a good candidate to capture the increase in death as time progresses as seen in our exploration of the data. The likelihood function for this model is given as:

$$\prod_{j=1}^{10}\prod_{i=1}^{n_j}\left(\frac{k_j}{\theta_j^{k_j}}\int_{t_{ij}-0.5}^{t_{ij}} s^{k_j-1}\exp\left(\frac{-s}{\theta_j}\right)^{k_j} ds\right)^{c_{ij}}\left(\exp\left(\frac{-t_{ij}}{\theta_j}\right)^{k_j}\right)^{1-c_{ij}}, \tag{4}$$

where $k > 0$ is the shape parameter and $\theta > 0$ is the scale parameter of the distribution.

Gamma Distribution Fixed Effects Model The gamma distribution arises naturally in many real-life phenomena and is a frequently used distribution for modeling non-negative random variables like time to death. Like Weibull, its hazard functions can be either monotonically increasing or decreasing, based on the value of the scale and shape parameters (Cox et al 2007). The likelihood function for the gamma model is given as:

$$\prod_{j=1}^{10}\prod_{i=1}^{n_j}\left(\frac{1}{\Gamma(k_j)\theta_j^{k_j}}\int_{t_{ij}-0.5}^{t_{ij}} s^{k_j-1}\exp\left(\frac{-s}{\theta_j}\right) ds\right)^{c_{ij}}\left(1-\gamma(\frac{t_{ij}}{\theta_j}, k_j)\right)^{1-c_{ij}}, \tag{5}$$

where

$$\gamma\left(\frac{t_{ij}}{\theta_j}, k_j\right) = \frac{1}{\Gamma(k_j)\theta_j^{k_j}}\int_0^{t_{ij}} s^{k_j-1}\exp\left(\frac{-s}{\theta_j}\right) ds.$$

Here $k > 0$ is the shape parameter and $\theta > 0$ is the scale parameter of the distribution. Note in the above expression, $\Gamma(k_j)$ is the gamma function evaluated at k_j while $\gamma\left(\frac{t_{ij}}{\theta_j}, k_j\right)$ defined above is known as the lower incomplete gamma function.

Lognormal Distribution Fixed Effects Model Unlike in the Weibull and Gamma cases, the hazard function of a log-normally distributed random variable need not be monotone. In fact, it can have an upward arc shape which makes it a good candidate for

our data to capture a possible peak in hazard during our study period. The likelihood function for this model is given as:

$$\prod_{j=1}^{10}\prod_{i=1}^{n_j}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\int_{t_{ij}-0.5}^{t_{ij}}\frac{1}{s}\exp\left(\frac{-(\log(s)-\mu_j)^2}{2\sigma_j^2}\right)ds\right)^{c_{ij}} \left(\frac{1}{2}-\frac{1}{2}\text{erf}\left(\frac{\log(t_{ij})-\mu_j}{\sqrt{2}\sigma_j}\right)\right)^{1-c_{ij}}, \tag{6}$$

where,

$$\text{erf}\left(\frac{\log(t_{ij})-\mu_j}{\sqrt{2}\sigma_j}\right)=\frac{2}{\sqrt{\pi}}\int_0^{\frac{\log(t_{ij})-\mu_j}{\sqrt{2}\sigma_j}}\exp(-s^2)ds.$$

Here, μ is the mean and σ is the standard deviation of the associated normal distribution. On a logarithmic scale, μ and σ are called the location parameter and the scale parameter, respectively.

3.2.2 Model with clustering

We next consider a set of more complex models which relax the assumption of genotype independence and allow for clusters depending on some known biological properties of the genotypes. For example, the diet group which did not feed on a soybean leaf would be expected to have a different distribution (and consequently, hazard shape) compared to groups that fed on soybean leaf disks. Figure 1 further supports this hypothesis. Additionally, as discussed earlier, results from Underwood et al (2000, 2002) can be used to cluster the genotypes as inducible and non-inducible. The non-inducible genotypes include Bragg, Gasoy and Tracy and inducible genotypes include Braxton, Clark, Cook, Davis, Stonewall and Williams. Thus, the likelihood for a three-cluster fixed-effect model can be expressed as:

$$\prod_{j=1}^{3}\prod_{i=1}^{n_j}\left(\int_{t_{ij}-0.5}^{t_{ij}} f(s\mid\Theta_j)\,ds\right)^{c_{ij}} S(t_{ij}\mid\Theta_j)^{1-c_{ij}}, \tag{7}$$

where j denotes the cluster ($j=1,2,3$ corresponding to inducible, non-inducible and diet), and i denotes the larvae within each cluster. Worms within each cluster are assumed to have the same hazard rate, depending on the cluster-specific set of parameters Θ_j. In principle, cluster time-to-events can belong to different families of distributions, although in our analysis we will only consider the same family (log-normal) for all clusters.

3.2.3 Random effects model

We next considered a set of random effects models, that presents a way to explicitly characterize the heterogeneity among the groups. These models are based on the premise that each group is allowed to have some unique distribution aspects while sharing some commonalities, described via the common random-effects distribution:

$$\prod_{j=1}^{10} g(\Theta_j \mid \Psi) \prod_{i=1}^{n_j} \left(\int_{t_{ij}-0.5}^{t_{ij}} f(s \mid \Theta_j)\, ds \right)^{c_{ij}} S(t_{ij} \mid \Theta_j)^{1-c_{ij}}. \tag{8}$$

Here j again denotes the genotype groups, $j = 1, \ldots, 10$. We further assume that each Θ_j is an independent and identically distributed draw from the common random effect distribution $g(\Theta_j \mid \Psi)$, depending on a common parameter Ψ which is also to be estimated from the data. The variance of this distribution is considered the heterogeneity parameter; the larger this variance, the less alike the groups are, and the higher the level of heterogeneity. We implement two different generalized gamma density functions, namely lognormal and gamma.

Gamma random effect model This model assumes that each group-specific parameter $\Theta_j = (k_j, \theta_j)$ is drawn independently from a gamma distribution with the common parameter $\Psi = (k_a, \theta_a, k_b, \theta_b)$, as follows:

$$\begin{aligned} k_j &\sim \mathcal{G}a(k_a, \theta_a), \\ \theta_j &\sim \mathcal{G}a(k_b, \theta_b), \end{aligned}$$

where the subscripts a and b denote the hyper-parameters associated with the gamma distributions for k_j and θ_j parameters respectively. The likelihood function for this model is given as:

$$\prod_{j=1}^{10} \left\{ \left(\frac{1}{\Gamma(k_a)\theta_a^{k_a}} (k_j)^{k_a-1} \exp\left(\frac{-k_j}{\theta_a} \right) \right) \left(\frac{1}{\Gamma(k_b)\theta_b^{k_b}} (\theta_j)^{k_b-1} \exp\left(\frac{-\theta_j}{\theta_b} \right) \right) \right.$$
$$\left. \prod_{i=1}^{n_j} \left(\frac{1}{\Gamma(k_j)\theta_j^{k_j}} \int_{t_{ij}-0.5}^{t_{ij}} s^{k_j-1} \exp\left(\frac{-s}{\theta_j} \right) ds \right)^{c_{ij}} \left(1 - \gamma(\frac{t_{ij}}{\theta_j}, k_j) \right)^{1-c_{ij}} \right\}. \tag{9}$$

Lognormal random effect model This model assumes that each group-specific parameter $\Theta_j = (\mu_j, \sigma_j)$ is independently drawn from a lognormal distribution with the common parameter $\Psi = (\mu_a, \sigma_a, \mu_b, \sigma_b)$, as follows:

$$\begin{aligned} \mu_j &\propto LN(\mu_a, \sigma_a), \\ \sigma_j &\propto LN(\mu_b, \sigma_b), \end{aligned}$$

where the subscripts a and b denote the hyper parameters associated with the lognormal distributions for μ_j and σ_j parameters respectively. The likelihood function for this

model is given as:

$$\prod_{j=1}^{10}\left\{\left(\frac{1}{\mu_j\sigma_a\sqrt{2\pi}}\exp\left(-\frac{(\log(\mu_j)-\mu_a)^2}{2\sigma_a^2}\right)\right)\right.$$
$$\times\left(\frac{1}{\sigma_j\sigma_b\sqrt{2\pi}}\exp\left(-\frac{(\log(\sigma_j)-\mu_b)^2}{2\sigma_b^2}\right)\right)$$
$$\times\prod_{i=1}^{n_j}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\int_{t_{ij}-0.5}^{t_{ij}}\frac{1}{s}\exp\left(\frac{-(\log(s)-\mu_j)^2}{2\sigma_j^2}\right)ds\right)^{c_{ij}}$$
$$\left.\times\left(\frac{1}{2}-\frac{1}{2}\text{erf}\left(\frac{\log(t_{ij})-\mu_j}{\sqrt{2}\sigma_j}\right)\right)^{1-c_{ij}}\right\}. \tag{10}$$

3.3 Semi-parametric analysis

In survival analysis, semi-parametric analysis refers to assuming a non-parametric form for the hazard (or survival) function and a parametric form for the covariate effects. A popular semi-parametric model is the Cox proportional hazard model (Cox 1972):

$$h(t) = h_{base}(t)\exp(X'\beta), \tag{11}$$

where β is a vector of covariate effects, $h_{base}(t)$ is the underlying baseline hazard function, and h is the hazard function for the group characterized by the unique combination of covariate levels X. Unlike in the parameter models where the $h_{base}(t)$ was specified using a known family of probability distributions, the Cox proportional hazard model does not assume a form for the baseline hazard and treats it as a nuisance parameter in estimation. However, for models where the primary interest lies in estimating the survival function (or equivalently, the hazard function; see Appendix 1), the baseline hazard is not a nuisance parameter. In order to estimate survival curves using Cox proportional hazard model, the Breslow estimator (Breslow 1972) is widely used to estimate the cumulative baseline hazard and subsequently the baseline survival function. It is given by,

$$\hat{\Lambda}_0(t) = \sum_{\tau_j<t}\left(\frac{d_j}{\sum_{k\in R(\tau_j)}\exp(\hat{\beta}'x_k)}\right), \tag{12}$$

where d_j is the number of deaths at time t_j and $R(\tau_j)$ is the risk set (number of survivors) at time t_j.The estimator directly uses the $\hat{\beta}$ parameter estimates from Cox's maximum partial likelihood estimator in estimating the hazard for a small time period. However, the estimator is known to be an inconsistent estimator of the hazard rate (Burr 1994).

3.3.1 Bayesian semi-parametric analysis

Here, we present a semi-parametric analysis in a Bayesian framework that estimates the hazard function explicitly. All the priors for our Bayesian models presented are chosen to be vague, in order for the results from these models to be comparable to the likelihood-based results from our parametric models.

The multiresolution hazard (MRH) model is a Bayesian semi-parametric model which enables us to estimate the baseline hazard rate $h_{base}(t)$ jointly with the covariate effects in the model. The approach consists of choosing a set of discretized time points $t_0, t_1, \ldots, t_J$ so that J represents the total number of bins across the time interval of interest. We set $J = 2^M$ ($M \in \mathbb{Z}^+$) to take advantage of the known and fast multiresolution wavelet methods (Bouman et al 2005) for estimation.

The MRH model is based on the partition of the baseline cumulative hazard into hazard increments as follows:

$$d_j = H_{base}(t_j) - H_{base}(t_{j-1}) = \int_{t_{j-1}}^{t_j} h_{base}(s)ds,$$

where $h_{base}(s)$ is the baseline hazard rate at time s. (Bouman et al 2005; Dukic and Dignam 2007). The hazard increments d_j can now be chosen such that the prior beliefs about the underlying hazard function can be incorporated into the model. We follow the notation introduced in Bouman et al (2005) where $H_{0,0}$ is the total cumulative baseline hazard over the entire time period $(0, t_J)$ (so $H_{0,0} = H(t_J)$), and $H_{M,i-1} = d_i$ for $i = 1, \ldots, J$. Then the multiresolution hazard tree is recursively defined by $H_{m-1,p} = H_{m,2p} + H_{m,2p+1}$ for $m = 0, \ldots, M$ and $p = 0, \ldots, 2^{m-1} - 1$. Here, m is the current level of resolution and p is the position within that level. Hence, the model splits the initial total cumulative hazard $H(t_J)$ into finer components with each additional level of resolution until we finally get to the bottom of the tree with the hazard increments d_j. The recursive splits of H over different branches is defined as $R_{m,p} = H_{m,2p}/H_{m-1,p}$, and are known as the 'split parameters'. Therefore, we can parametrize the hazard increments by $H_{0,0}$ and the "splits" by $R_{1,0}, \ldots, R_{M,2^{M-1}-1}$ (denoted as $R_{m,p}$ hereafter). Figure 2 demonstrates how the MRH approximates the baseline hazard shape. The complete hazard rate prior is specified by putting priors on all the aforementioned tree parameters: a Gamma prior is placed on H, and Beta priors on each split parameters $R_{m,p}$:

$$\begin{aligned} H &\sim \mathcal{G}a(a, \lambda), \\ R_{m,p} &\sim \mathcal{B}e(2\gamma_{m,p}k^m a, 2(1-\gamma_{m,p})k^m a). \end{aligned} \tag{13}$$

Dukic and Dignam (2007) extended the MRH model to include a hierarchical structure and relax the proportional hazards assumption. One advantage of this method is that it permits a different baseline hazard shapes according to the group, and clustering of baseline hazards within each group.

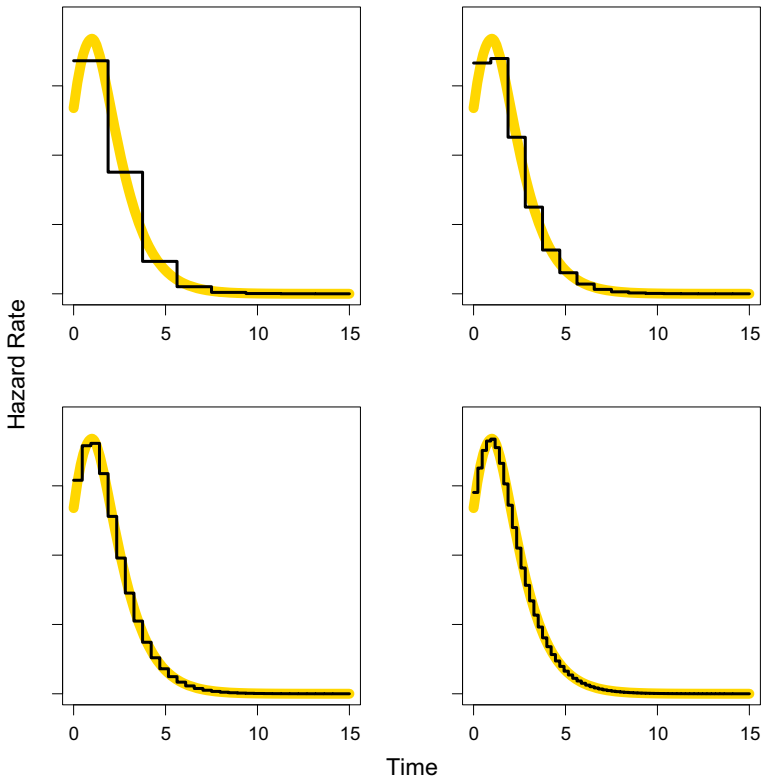


Fig. 2 This figure shows the MRH approximation of the hazard function. The yellow curve is the true hazard function, and the black lines are MRH approximations which get progressively better with finer time resolutions (shown for 4 different resolution values of M: 3, 4, 5 and 6)

3.3.2 Candidate MRH models

For larva i in genotype (stratum) s, with failure (or censoring) time at $T_{i,s} \in [0, t_J)$, its likelihood contribution is:

$$L_{i,s}(\boldsymbol{\beta} \mid T_{i,s} X_{i,s}) = \left(h_{0,s}(T_{i,s}) \exp(X'_{i,s}\boldsymbol{\beta})\right)^{c_i} S_{0,s}(T_{i,s})^{\exp(X'_{i,s}\boldsymbol{\beta})}, \tag{14}$$

where $h_{0,s}$ denotes the baseline hazard rate for genotype (strata) s, $X_{i,s}$ is the larval covariate vector and $S_{0,s}$ denotes the baseline survival function for the genotype s. Additionally, the c_i is the censoring variable which is 1 if larva i had an observed death and 0 if it was right-censored.

1. **Proportional Hazard Assumption**
   Under the proportional hazard assumption, we treat genotype and diet groups as

covariates, so the log-likelihood for all N larvae is:

$$\log L(\mathbf{T}|\boldsymbol{\beta}, \mathbf{H}, \mathbf{R_{m,p}}, \mathbf{X}) = \sum_{i=1}^{N} c_i\{\log(h_0(T_i)) + X_i'\boldsymbol{\beta}\} - \exp(X_i'\boldsymbol{\beta})H_0(T_i)\}, \quad (15)$$

where $H_0(T) = -\log S_0(T)$, is the cumulative baseline hazard. Here, X' represents the $N \times 10$ design matrix of 10 group indicator covariates. The columns are binary coded representing whether a larva, represented by a row, belongs to the ith group.

2. **Non-proportional Hazard Assumption**
The log-likelihood for all N larvae in $\mathscr{S}$ strata together is:

$$\log L(\mathbf{T}|\boldsymbol{\beta}, \mathbf{H}, \mathbf{R_{m,p,s}}, \mathbf{X}) = \sum_{s=1}^{\mathscr{S}} \sum_{i=1}^{N_s} c_i\{\log(h_{0,s}(T_{i,s})) - H_{0,s}(T_{i,s})\}. \quad (16)$$

Here, we have 10 strata representing nine genotype and one diet group. In this model, we have no covariates so the design matrix and β are no longer present in the likelihood function. Instead, we have the genotypes and diet being categorized as strata s with their own baseline hazards. Thus, each N_s represents the number of larvae in genotype s.

3. **Non-proportional Hazard Assumption on Genotype Clusters**
We can look at MRH models where we impose a similar grouping structure as in our parametric analysis discussed in Sect. 3.2.2. The genotypes can be grouped into clusters of inducible and non-inducible defenses. Similarly, the diet is its own group due to the lack of soybean intake by the fall armyworm. We looked at two different grouping structures:

   (a) Soybean versus non-soybean diet: The log-likelihood for all N larvae in 2 strata will be:

$$\log L(\mathbf{T}|\boldsymbol{\beta}, \mathbf{H}, \mathbf{R_{m,p,s}}, \mathbf{X}) = \sum_{s=1}^{2} \sum_{i=1}^{N_s} c_i\{\log(h_{0,s}(T_{i,s})) - H_{0,s}(T_{i,s})\}. \quad (17)$$

   (b) Inducible versus non-inducible versus non-soybean diet: We will have a similar log-likelihood for all N larvae in $\mathscr{S}$ strata where we have 3 strata:

$$\log L(\mathbf{T}|\boldsymbol{\beta}, \mathbf{H}, \mathbf{R_{m,p,s}}, \mathbf{X}) = \sum_{s=1}^{3} \sum_{i=1}^{N_s} c_i\{\log(h_{0,s}(T_{i,s})) - H_{0,s}(T_{i,s})\}. \quad (18)$$

   The corresponding strata in this model are non-inducible, inducible, and the diet group.

3.4 Model fitting

The parametric models were fitted using a custom maximum likelihood optimization code in MATLAB, based on Newton-Raphson algorithm with numerical first- and second-order derivatives. We specified the starting values based on the group sample statistics such as the mean and variance. The MRH models were estimated using the Bayesian framework and MCMC in the R package 'MRH' (Dukic and Dignam 2007; Bouman et al 2005; Hagar et al 2014; Chen et al 2014) where $M = 5$ was chosen so that there were $J = 2^M = 32$ bins where each bin represents a 12 h time interval. The MRH package assessed convergence through standard graphical techniques and Gelman-Rubin diagnostics.

3.5 Comparison of candidate models

We examine several information criteria in order to compare the candidate models. Measures of predictive accuracy are typically based on the deviance (the log predictive density of the data given a point estimate of the fitted model, multiplied by -2; i.e., $-2 \log p(y|\hat{\theta})$). When comparing models of different complexity, a penalty based on that complexity is usually added, reflecting the fact that larger more complex models can generally fit the data better. The information criteria described below impose penalties based on the number of parameters in a model, which is the most common and simplest approximation of model complexity. Using such criteria for hierarchical and Bayesian models is not straight forward, as the number of parameters is not a well defined concept in these models due to often strong dependencies between parameters themselves, even though some catered solutions have been proposed (Donohue et al 2011; Vaida and Blanchard 2005; Greven and Kneib 2010).

The first information criterion we consider is the Akaike Information Criterion (AIC) as defined by Akaike (1973):

$$AIC = -2 \log p(y|\hat{\theta}_{mle}) + 2k, \tag{19}$$

where k is the number of parameters estimated in the model. Additionally, we also used the Bayesian Information Criterion (BIC), where

$$BIC = -2 \log p(y|\hat{\theta}_{mle}) + k \log(n), \tag{20}$$

where k is again the number of parameters estimated in the model. The BIC provides a higher penalty for larger sample size, n, compared to AIC, and hence favors simpler models (Gelman et al 2013). For our Bayesian models, we also calculated the Watanabe–Akaike or widely available information criterion (WAIC) as defined in Gelman et al (2013):

$$WAIC = -2 \text{ lppd} + 2p_{WAIC}, \tag{21}$$

where p_{WAIC} is defined as the effective number of parameters,

$$p_{WAIC} = \sum_{i=1}^{n} \text{var}_{post}(\log p(y_i|\theta)). \tag{22}$$

Similarly, the log pointwise predictive density (lppd) in Eq. 21 is calculated as:

$$\text{computed lppd} = \sum_{i=1}^{n} \log \left(\frac{1}{M} \sum_{m=1}^{M} p(y_i|\theta^m) \right) \tag{23}$$

over M posterior draws and n data points. By comparing across all the information criteria, we hope to choose the best model that accounts for both the goodness of fit and the number of fitted parameters.

4 Results

For each candidate model (Table 2), we estimated the model parameters using either maximum likelihood or Bayesian posterior mean estimates (under vague priors), and calculated each of the information criteria scores. For the BIC and AIC computation in MRH models we used the total number of parameters (including prior parameters which are estimated from the data), because it is the worst case scenario for the Bayesian MRH models, as it puts the maximum amount of penalty possible on them. This avoids the difficulty with counting the effective number of parameters in Bayesian hierarchical models (Spiegelhalter et al 2002), by giving an upper bound on the BIC and AIC for these models. For the random effect models, we presented a range for AIC and BIC, corresponding to the lowest penalty (counting only the likelihood-level parameters and not the random effect distribution parameters) and the highest penalty (counting all parameters, including those from the random effect distribution.)

In general, the Bayesian semi-parametric models had uniformly lower model selection criteria than any of the parametric models. The Bayesian models performed better under the worst penalty than the parametric models under the lowest penalty. Within the parametric models, lognormal distribution models performed better than the rest. The best-fit model across all criteria was the MRH model with the proportional hazard assumption (Model 8).

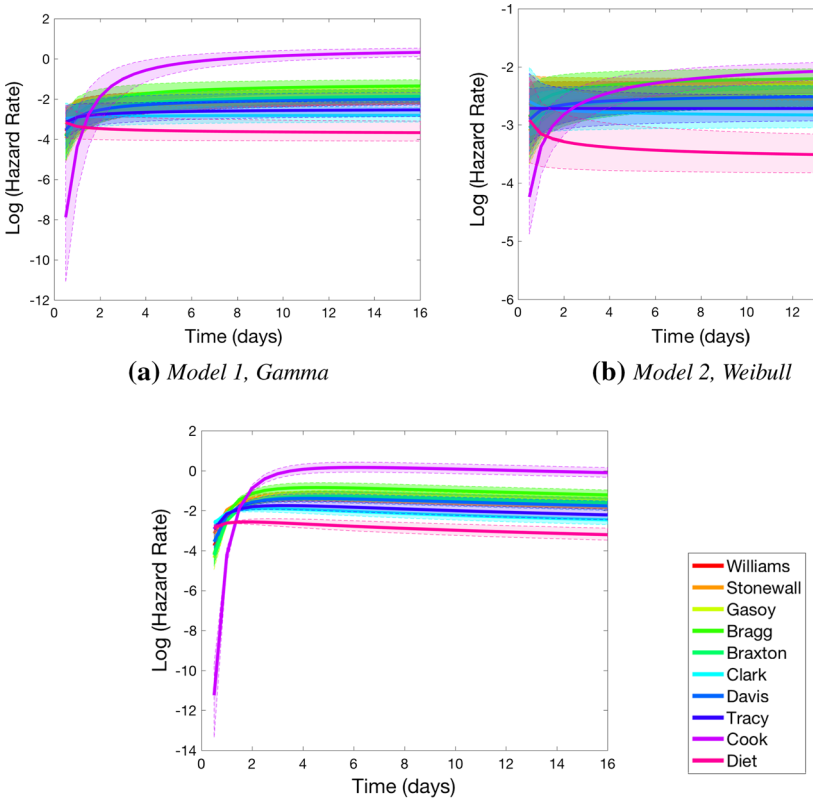
In terms of whether clustering better explained the data, clustering the groups based on plant induction type did not seem to produce a better model. The clustered models in the parametric family (Models 4 and 5) did not perform much better than the non-clustered lognormal (Model 3). Similarly, the clustered MRH models (Models 10 and 11) did not perform better than our best-fit model which assumes proportional hazard for individual genotypes (i.e., no clusters).

Figure 3 shows the estimated log hazard functions for our parametric models using gamma, Weibull and lognormal distributions. The Cook genotype had a higher hazard estimate than the rest of the genotypes in two of the three models. All of the parametric

Table 2 Information criteria Δ_{IC} (IC - min(IC)) calculated for each of the models used to fit the larval time to death data

Model (summary)	k	Δ BIC	Δ AIC	p_{WAIC}	Δ WAIC
1. Individual Gamma	20	1553.2	1453.9	–	–
2. Individual Weibull	20	1606.4	1507.1	–	–
3. Individual Lognormal	20	1363.9	1264.6	–	–
4. Two-cluster Lognormal	4	1406.2	1237.7	–	–
5. Three-cluster Lognormal	6	1410.1	1250.4	–	–
6. Gamma RE	(20, 24)	(1626.4, 1634.4)	(1527.1, 1552.4)	–	–
7. Lognormal RE	(20, 24)	(1387.6, 1395.6)	(1288.3, 1313.6)	–	–
8. MRH PH	**43**	**0**	**0**	**30.5**	**0**
9. MRH NPH	340	1170.3	2453.0	205.8	458.3
10. MRH NPH 2 clusters (a)	68	69.3	177.2	25.3	20.1
11. MRH NPH 3 clusters (b)	102	152.7	407.6	40.0	36.7

Models 1 through 7 are the parametric models and Models 8 through 11 are the Bayesian semi-parametric models. The model in bold provides the best fit (the lowest scores overall). k represents the number of parameters in the model and p_{WAIC} represents the estimated effective number of parameters in the Bayesian models. The sample size (n) was 555. The range of k for Models (6) and (7) corresponds to counting only population level parameters (the smallest penalty) and counting all parameters including the hyperparameters (the highest penalty). The Bayesian MRH models' k was computed by counting all parameters as well, reflecting the highest penalty possible for those models. Labels (a) and (b) in titles of models 10 and 11 refer to Eqs. (17) and (18), respectively



(a) *Model 1, Gamma*

(b) *Model 2, Weibull*

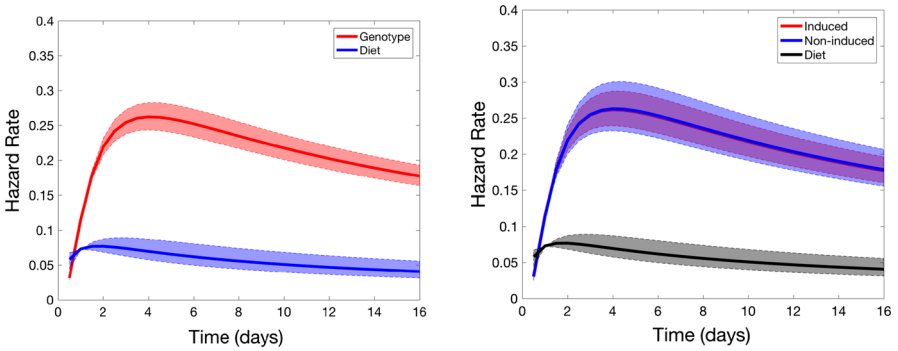
(c) *Model 3, Lognormal*

Fig. 3 Log hazard rates resulting from the maximum likelihood estimation of the individual parametric models for each genotype and diet. The shaded regions represents the 95% pointwise confidence interval associated with each estimate

models seem to have a difficult time capturing the decrease in the hazard function over time, although the lognormal model does better than the rest.

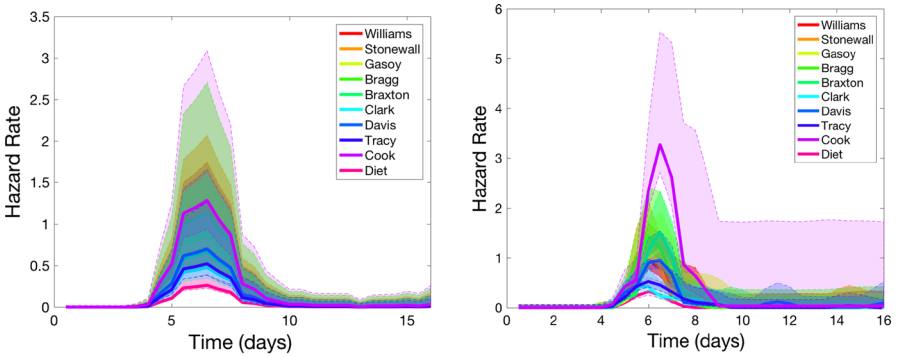
Figure 4 shows the estimated hazard functions for our clustered parametric models with lognormal distributions. The two cluster model in Fig. 4a shows a much higher hazard rate (approximately 5 times higher) associated with consuming infected soybean leaves compared to the diet group that did not consume soybean leaves. We see a similar hazard behavior in the three cluster model in Fig. 4b. Interestingly, the almost entirely overlapping hazard curves for the induced and non-induced genotypic groups imply that there is very little difference in hazard associated with the two groups except for a wider confidence band associated with the non-induced genotypes. However, as in the fixed effects individual model, the cluster models are unable to entirely capture the rapid decrease in hazard over time.

Figures 5 and 6 represent the hazard functions estimated by the MRH models. We see a similar characterization with Cook having the highest hazards. We also see similar hazard differences between the clusters and diet group in MRH and MLE



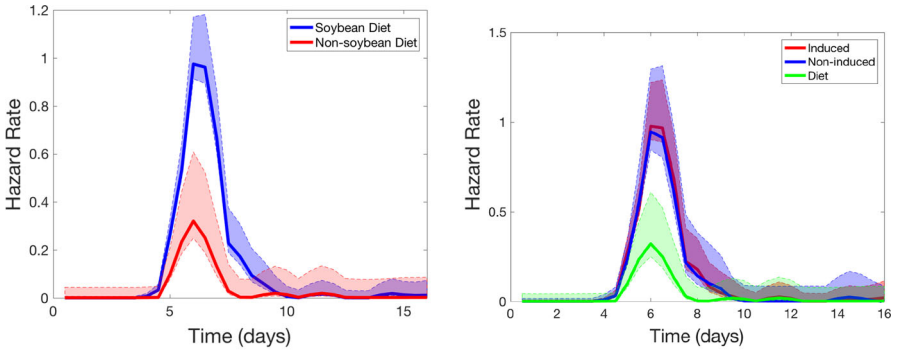
(a) *Model 4, two clusters* **(b)** *Model 5, three clusters*

Fig. 4 Hazard rates resulting from the maximum likelihood estimation for parametric analysis of clustered models. The shaded regions represents the 95% pointwise confidence interval associated with each estimate



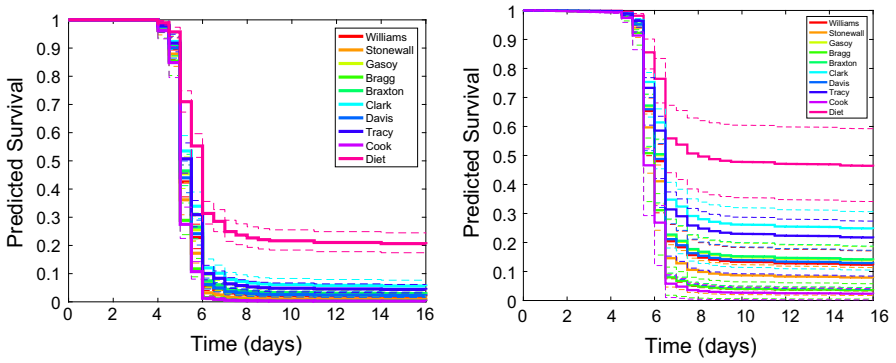
(a) *Model 8, Proportional Hazard* **(b)** *Model 9, Non-proportional Hazard*

Fig. 5 Hazard rates for MRH models under the proportional hazard and the non-proportional assumptions. The shaded regions represents the 95% pointwise credible interval associated with each estimate



(a) *Model 10, Two clusters under NPH* **(b)** *Model 11, Three clusters under NPH*

Fig. 6 Hazard rates of clustered MRH models under the non-proportional assumption. The shaded regions represents the 95% pointwise credible interval associated with each estimate



(a) *Cox Proportional Hazard* **(b)** *Model 8, MRH Proportional Hazard*

Fig. 7 A comparison of the predicted survival curves from our best-fit model and the semi-parametric Cox proportional hazard model

models, seen in Figs. 4 and 6. The cluster MRH NPH models in Fig. 6 also show a higher hazard associated with being on the soybean diet. As in the lognormal three cluster model in Fig. 4b, the hazard curves for the induced and non-induced genotype are very similar in the MRH three cluster model in Fig. 6b. Thus, both our parametric cluster model and the MRH cluster model show negligible difference in the hazard associated with the two types of genotypic cluster. However, the MRH models capture the decrease in hazard over time much better, which is an important feature of the data. The hazard rate for the fall armyworm seems to peak around Day 7.5 after the start of the infection and decays close to 0 as time passes. This hazard characteristic is seen in all the MRH models.

Figure 7 presents a visual comparison of the estimated survival curves from our best fit model, Model 8 and the Cox proportional hazard model given in Eq. (11). The baseline survival for the Cox model was estimated using the Breslow estimator given in Eq. (12). The estimated survival from the Cox model were much lower than those from our best fit MRH model in all groups, particularly around the time that the hazard peaks in Figs. 5 and 6. Moreover, the MRH-based model is able to provide a range of survival curves for different genotypes and better represent the differences among those, compared to the Cox proportional model which predicts low survival (< 0.1) after Day 6 for most of the genotypes.

5 Discussion

Our best-fit model assumes a proportional hazard assumption under the Bayesian semi-parametric framework of MRH. Given the severe penalties of BIC, AIC and WAIC for this relatively complex model, it is clear that a multiresolution approach to characterizing the hazard allows a researcher to draw the maximum amount of information from these data. The parametric models appear to capture the initial increase in hazard well but fail at accurately modeling the decrease. The widely used Weibull and gamma models cannot capture any decrease in hazard, as seen in Fig. 3. This is

because while both the Weibull and gamma models are flexible and allow for hazard rates that are non-constant, they are constrained to have monotonic hazard shapes. Hence, they will fit any dataset with a non-monotonic hazard rate poorly. However, the lognormal distribution does allow for non-monotonic shapes of hazard function (e.g., inverse bath-tub shaped) but still fails to capture the decrease in the hazard function rapidly enough due to the strong smoothness property. These approaches simply fail to maximize the information drawn from the mortality data in this case. Hence, our results demonstrate the pitfalls of assuming a commonly used parametric form for the mortality time for computational and analytic simplicity without further analysis of the hazard's shape. The apparent decrease in hazard modeled by all the MRH models suggests that the probability of dying for the larvae decreases dramatically after the hazard has peaked around day 7.5 of the infection. Biologically, this fits well with the fact that if the larvae are able to survive the infection up until a certain time point, they should be expected to survive well beyond that time point as well.

Figure 7 shows the estimated survival curves for different genotypes based on our semi-parametric MRH approach (where the baseline hazard is jointly estimated along with the covariate effects), and based on the Cox proportional hazard model (where the Breslow estimator (Breslow 1972) is used to estimate the baseline cumulative hazard). The Breslow estimator is a non-parametric maximum likelihood estimator for the cumulative baseline hazard estimate, and is based, in part, on the Cox partial likelihood covariate effect estimates. While the performance of this combination of non-parametric likelihood and partial likelihood estimators in finite samples is not entirely understood, it has been observed that it can lead to non negligible bias and underestimated uncertainty for the hazard function (Hagar and Dukic 2015). On the other hand, the Bayesian MRH model is explicitly formulated to estimate the joint finite-sample uncertainty through the joint posterior distribution, which is based on the joint likelihood for the hazard function and covariate effects. Hagar and Dukic (2015) present an extensive comparison of the performance of the MRH model with other commonly used, comparable semi-parametric survival models including the Cox model. They found that the Cox model based estimators for the baseline hazard function did not perform well in terms of bias and mean square error, unlike the MRH. Therefore, if accurate hazard shapes and a proper quantification of the associated uncertainty are of interest, including the option of relaxing the proportional hazard assumption, the MRH model is a valuable option.

Our results also show that there is clearly a larger hazard associated with consuming the virus with a soybean leaf compared to consuming the virus without ingesting leaf tissue. Our best-supported model treats the diet as the baseline and characterizes the hazards of each genotype as a proportion of this baseline. Thus, studying the covariate effect of the different genotypes helps us understand the risks or benefits associated with the fall armyworm food quality and its effect on the speed of kill. Given that the infection process in the field involves a tritrophic interaction that includes the host's food resources, our results show that ignoring the effect of the resource could be costly. Clearly, the data demonstrate that different soybean genotypes play a role in the time to death for the host. There has been extensive work done on host and pathogen variability and genetic diversity as a way of understanding host-pathogen interactions (Elderd et al 2008; Myers and Cory 2016; Kennedy et al 2014; Dwyer

et al 1997). However, the impact of plant genetic diversity on within host processes has not been studied or modeled extensively (but see Shikano et al 2017). Our results strongly suggest that any kind of within-host interaction model for the fall armyworm should include plant genotypic variability. The importance of the resources consumed by a host in helping or hindering individuals fighting off an infection should not be limited to just the interactions considered here (Lively et al 2014). The next step would be to move beyond the within-host effects and incorporate our findings in a model examining the population-level consequences of this tritrophic interaction that includes a plant/resource genotype component.

Interestingly, our clustering approach based on the level of induced resistance in the genotypes did not result in a better model for these data. While the results in Figs. 4 and 6 are striking, this still implies that induction grouping does not seem to tell the entire story in this case. In our experiments, undamaged leaf tissue was provided to the larvae as opposed to damaged leaf tissue, which may trigger the induction of additional defenses (Underwood et al 2002). Thus, some of the genotypes could have low inducible resistance but still have high constitutive defenses. Previous researchers have found no evidence of a correlation between induced and constitutive resistance (Underwood et al 2000). A clustering method based on constitutive defense would make for a good comparison but we did not examine how constitutive defenses affect the within-host process because the level of constitutive defenses in each of the genotypes was not available to us.

Through the comparison between the parametric approach and Bayesian semi-parametric approach, we have shown that standard methods may not be able to correctly characterize the hazard shapes and that the flexibility of the Bayesian semi-parametric approach does. When the data are censored, which is often the case in nature, the flexibility of the Bayesian semi-parametric approach may become even more important.

While we focused on plant genotypic variability, we have not considered other factors that can influence within-host processes. These include host genetic variability (Dwyer et al 1997; Elderd et al 2008; Páez et al 2015) and pathogen variability (Fleming-Davies et al 2015). Similarly, population-level processes could have effects on the within-host level process which we have not considered here. However, it would be a natural extension to our models and would be a reasonable avenue for future research. Specifically, our results highlight the importance of considering the tritrophic interaction between the host, its food resources, and its pathogen for disease driven systems.

We demonstrated the flexibility and utility of the MRH model in modeling time to death for interval- and right-censored data. However, the method can be also used for a variety of censored survival data, including those arising from mark-recapture studies where it is common to have missing or incomplete records. Specifically, if it is important to accurately model the hazard shape for use in population models, the MRH method can be easily implemented on left-, right-, or interval-censored data. In general, researchers have found the use of individual hazard models to be a powerful approach for analyzing mortality rates whether or not the data are censored (e.g., Zens and Peart 2003). Additionally, the flexibility of being able to estimate a combination of multiple proportional and non-proportional covariate effects in survival data makes this a useful tool in modeling most kinds of censored as well as uncensored data.

Lastly, these methods are applicable to any kind of non-death event data in ecology and evolution where hazard (albeit interpreted differently) is of primary interest, and show that flexibility of a semi-parametric approach can allow researchers to maximize the amount of information drawn from their data.

Acknowledgements We thank the Elderd Lab at Louisiana State University for their help and guidance with the experiments. We also thank Dr. Yolanda Hagar for her help with the Multiresolution Hazard (MRH) package. This work was funded by National Science Foundation (NSF) Grant 1316334 as part of the joint NSF-National Institutes of Health-USDA Ecology and Evolution of Infectious Diseases program. We would also like to thank the associate editor and reviewers for their insightful comments and suggestions.

Appendix 1: Survival and hazard functions

Survival function Let T be a non-negative random variable representing the waiting time until the occurrence of the event in question. Then, the cumulative distribution function, $F(t) = P(T < t)$, is the probability that the event has occurred by duration t. Hence, the survival function is the probability of survival beyond time t. It is given by

$$S(t) = P(T \geq t) = 1 - F(t). \tag{24}$$

It is a non-increasing function that starts at 1 and asymptotically goes to 0 as time goes to infinity.

Hazard function The hazard function can be defined as the probability of failure in an infinitesimally small period between t and $t + ct$ given the subject has survived until time t. In other words, it is an instantaneous rate of failure. Let $f(t)$ be the probability density of failure in an infinitesimally small period between t and $t + ct$. Then, the hazard function is defined as,

$$h(t) = \frac{f(t)}{S(t)}, \tag{25}$$

which is the probability that the event occurred in the interval t to ct divided by the probability that an individual survived to time t. It is also a non-negative function but unlike the survival function, it can take any shape and is not bounded between 0 and 1.

Cumulative hazard function The cumulative hazard function,

$$H(t) = \int_0^t h(y)dy$$

is also a measure of risk such that the higher the cumulative hazard, the greater the risk of failure by time t.

Substituting Eq. 24 into Eq. 25 we get,

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

Integrating both sides from 0 to t,

$$\begin{aligned}\int_0^t h(y)dy &= \int_0^t \frac{f(y)}{1-F(y)}dy\\ H(t) &= -\ln(1-F(t)) \quad [\text{Since, } F'(t) = f(t)]\\ H(t) &= -\ln(S(t))\\ \exp(-H(t)) &= S(t).\end{aligned}$$

Appendix 2: Associated Log-likelihood functions

– The corresponding log-likelihood function for Eq. 4,

$$\sum_{j=1}^{10}\sum_{i=1}^{n_j} c_{ij}\left[\log(k_j) - k_j\log(\theta_j) + \log\left(\int_{t_{ij}-0.5}^{t_{ij}} s^{k_j-1}\exp\left(\frac{-s}{\theta_j}\right)^{k_j} ds\right)\right] - (1-c_{ij})\left(\frac{t_{ij}}{\theta_j}\right)^{k_j}. \tag{26}$$

– The corresponding log-likelihood function for Eq. 5,

$$\sum_{j=1}^{10}\sum_{i=1}^{n_j} c_{ij}\log\left(\gamma\left(\frac{t_{ij}}{\theta_j}, k_j\right) - \gamma\left(\frac{t_{ij}-0.5}{\theta_j}, k_j\right)\right) + (1-c_{ij})\log\left(1-\gamma\left(\frac{t_{ij}}{\theta_j}, k_j\right)\right). \tag{27}$$

– The corresponding log-likelihood function for Eq. 6,

$$\sum_{j=1}^{10}\sum_{i=1}^{n_j} c_{ij}\log\left(\frac{1}{2}\text{erf}\left(\frac{\log(t_{ij})-\mu_j}{\sqrt{2}\sigma_j}\right) - \frac{1}{2}\text{erf}\left(\frac{\log(t_{ij}-0.5)-\mu_j}{\sqrt{2}\sigma_j}\right)\right) + (1-c_{ij})\log\left(\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\log(t_{ij})-\mu_j}{\sqrt{2}\sigma_j}\right)\right). \tag{28}$$

– The corresponding log-likelihood function for Eq. 9,

$$\sum_{j=1}^{10}\Bigg\{(k_a-1)\log(k_j)-\frac{k_j}{\theta_a}-\log(\Gamma(k_a))-k_a\log(\theta_a)+(k_b-1)\log(\theta_j)$$
$$-\frac{\theta_j}{\theta_b}-\log(\Gamma(k_b))-k_b\log(\theta_b)$$
$$+\sum_{i=1}^{n_j}c_{ij}\log\left(\gamma\left(\frac{t_{ij}}{\theta_j},k_j\right)-\gamma\left(\frac{t_{ij}-0.5}{\theta_j},k_j\right)\right)$$
$$+(1-c_{ij})\log\left(1-\gamma\left(\frac{t_{ij}}{\theta_j},k_j\right)\right)\Bigg\}. \tag{29}$$

– The corresponding log-likelihood function for Eq. 10,

$$\sum_{j=1}^{10}\Bigg\{-(\log(\mu_j)+\log(\sigma_a))-\frac{(\log(\mu_j)-\mu_a)^2}{2\sigma_a^2}-(\log(\sigma_j)$$
$$+\log(\sigma_b))-\frac{(\log(\sigma_j)-\mu_b)^2}{2\sigma_b^2}$$
$$+\sum_{i=1}^{n_j}c_{ij}\log\left(\frac{1}{2}\mathrm{erf}\left(\frac{\log(t_{ij})-\mu_j}{\sqrt{2}\sigma_j}\right)-\frac{1}{2}\mathrm{erf}\left(\frac{\log(t_{ij}-0.5)-\mu_j}{\sqrt{2}\sigma_j}\right)\right)$$
$$+(1-c_{ij})\log\left(\frac{1}{2}-\frac{1}{2}\mathrm{erf}\left(\frac{\log(t_{ij})-\mu_j}{\sqrt{2}\sigma_j}\right)\right)\Bigg\}. \tag{30}$$

References

Anderson RM, May RM (1981) The population dynamics of microparasites and their invertebrate hosts. Philos Trans R Soc Lond B Biol Sci 291(1054):451–524. https://doi.org/10.1098/rstb.1981.0005

Bi JL, Felton GW (1995) Foliar oxidative stress and insect herbivory: primary compounds, secondary metabolites, and reactive oxygen species as components of induced resistance. J Chem Ecol 21(10):1511–1530. https://doi.org/10.1007/BF02035149

Botella MA, Xu Y, Prabha TN, Zhao Y, Narasimhan ML, Wilson KA, Nielsen SS, Bressan RA, Hasegawa PM (1996) Differential expression of soybean cysteine proteinase inhibitor genes during development and in response to wounding and methyl jasmonate. Plant Physiol 112(3):1201–1210

Bouman P, Dukic V, Meng XL (2005) A Bayesian multiresolution hazard model with application to an AIDS reporting delay study. Stat Sin 15(2):325–357

Breslow NE (1972) Contribution to discussion of paper by Dr. Cox. J R Stat Assoc B 34:216–217

Brookmeyer R (2014) Median survival time. Wiley, New York. https://doi.org/10.1002/9781118445112.stat06043

Burr D (1994) On inconsistency of Breslow's estimator as an estimator of the hazard rate in the Cox model. Biometrics 50(4):1142–1145. http://www.jstor.org/stable/2533450

Chen Y, Hagar Y, Dignam J, Dukic V (2014) Pruned multiresolution hazard (pmrh) models for time-to-event data. Bayesian Anal (in review)

Cory JS, Hoover K (2006) Plant-mediated effects in insect-pathogen interactions. Trends Ecol Evol 21(5):278–286. https://doi.org/10.1016/j.tree.2006.02.005

Cory JS, Myers JH (2003) The ecology and evolution of insect baculoviruses. Annu Rev Ecol Evol Syst 34(1):239–272. https://doi.org/10.1146/annurev.ecolsys.34.011802.132402
Cox DR (1972) Regression models and life-tables. J R Stat Soc Ser B (Methodol) 34(2):187–220
Cox DR, Oakes D (1984) Analysis of survival data. CRC Press, Boca Raton
Cox C, Chu H, Schneider MF, Muoz A (2007) Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Stat Med 26(23):4352–4374. https://doi.org/10.1002/sim.2836
Dobson AJ (2002) An introduction to generalized linear models, 2nd edn. Chapman & Hall/CRC texts in statistical science series. Chapman & Hall/CRC, Boca Raton
Donohue M, Overholser R, Xu R, Vaida F (2011) Conditional akaike information under generalized linear and proportional hazards mixed models. Biometrika 98(3):685–700
Dukic V, Dignam J (2007) Bayesian hierarchical multiresolution hazard model for the study of time-dependent failure patterns in early stage breast cancer. Bayesian Anal 2(3):591–609. https://doi.org/10.1214/07-BA223
Dwyer G, Elkinton JS, Buonaccorsi JP (1997) Host heterogeneity in susceptibility and disease dynamics: tests of a mathematical model. Am Nat 150(6):685–707. https://doi.org/10.1086/286089
Dwyer G, Dushoff J, Elkinton JS, Levin SA (2000) Pathogen-driven outbreaks in forest defoliators revisited: building models from experimental data. Am Nat 156(2):105–120. https://doi.org/10.1086/303379
Elderd BD, Reilly J (2014) Warmer temperatures increase disease transmission and outbreak intensity in a host-pathogen system. J Anim Ecol 83:838–849
Elderd BD, Dushoff J, Dwyer G (2008) Host-pathogen interactions, insect outbreaks, and natural selection for disease resistance. Am Nat 172(6):829–842. https://doi.org/10.1086/592403
Elderd BD, Rehill BJ, Haynes KJ, Dwyer G (2013) Induced plant defenses, host-pathogen interactions, and forest insect outbreaks. Proc Natl Acad Sci 110(37):14,978–14,983. https://doi.org/10.1073/pnas.1300759110
Elkinton JS, Liebhold AM (1990) Population dynamics of gypsy moth in North America. Annu Rev Entomol 35(1):571–596. https://doi.org/10.1146/annurev.en.35.010190.003035
Farrar RR, Ridgway RL (2000) Host plant effects on the activity of selected nuclear polyhedrosis viruses against the corn earworm and beet armyworm (Lepidoptera: Noctuidae). Environ Entomol 29(1):108–115. https://doi.org/10.1603/0046-225X-29.1.108
Fleming-Davies AE, Dukic V, Andreasen V, Dwyer G (2015) Effects of host heterogeneity on pathogen diversity and evolution. Ecol Lett 18(11):1252–1261
Foster MA, Schultz JC, Hunter MD (1992) Modelling gypsy moth-virus-leaf chemistry interactions: implications of plant quality for pest and pathogen dynamics. J Anim Ecol 61(3):509–520. https://doi.org/10.2307/5606
Fuxa JR (1982) Prevalence of viral infections in populations of fall armyworm, Spodoptera frugiperda, Southeastern Louisiana. Environ Entomol 11(1):239–242
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman and Hall/CRC, Boca Raton
Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. Stat Comput 24(6):997–1016
Greenwood M (1926) A report on the natural duration of cancer. Reports on Public Health and Medical Subjects Ministry of Health (33)
Greven S, Kneib T (2010) On the behaviour of marginal and conditional aic in linear mixed models. Biometrika 97(4):773–789
Hagar Y, Albers D, Pivovarov R, Chase H, Dukic V, Elhadad N (2014) Survival analysis with electronic health record data: experiments with chronic kidney disease. Stat Anal Data Min ASA Data Sci J 7(5):385–403. https://doi.org/10.1002/sam.11236
Hagar Y, Dignam JJ, Dukic V (2017) Flexible modeling of the hazard rate and treatment effects in long-term survival studies. Stat Methods Med Res 26(5):2455–2480
Hagar Y, Dukic V (2015) Comparison of hazard rate estimation in R. arXiv:1509.03253
Hinds W, Dew J (1915) The grass worm or fall army worm. Techical Report, Bulletin no. 186, Alabama Agricultural Experiment Station
Hoover K, Kishida KT, DiGiorgio LA, Workman J, Alaniz SA, Hammock BD, Duffey SS (1998) Inhibition of baculoviral disease by plant-mediated peroxidase activity and free radical generation. J Chem Ecol 24(12):1949–2001. https://doi.org/10.1023/A:1020777407980

Hudson PJ, Dobson AP, Newborn D (1998) Prevention of population cycles by parasite removal. Science 282(5397):2256–2258

Ibrahim Ali M, Young SY, Felton GW, McNew RW (2002) Influence of the host plant on occluded virus production and lethal infectivity of a baculovirus. J Invertebr Pathol 81(3):158–165. https://doi.org/10.1016/S0022-2011(02)00193-3

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53(282):457. https://doi.org/10.2307/2281868

Keating ST, Yendol WG (1987) Influence of selected host plants on gypsy moth (Lepidoptera: Lymantriidae) larval mortality caused by a baculovirus. Environ Entomol 16(2):459–462. https://doi.org/10.1093/ee/16.2.459

Kennedy DA, Dukic V, Dwyer G (2014) Pathogen growth in insect hosts: Inferring the importance of different mechanisms using stochastic models and response-time data. Am Nat 184(3):407–423

Kleinbaum DG, Klein M (2006) Survival analysis: a self-learning text. Springer Science & Business Media, New York

Lively CM, de Roode JC, Duffy MA, Graham AL, Koskella B, Day ET (2014) Interesting open questions in disease ecology and evolution. Am Nat 184(S1):S1–S8. https://doi.org/10.1086/677032

Miller LK (1997) Baculoviruses. Kluwer Academic, New York

Muenchow G (1986) Ecological use of failure time analysis. Ecology 67(1):246–250. https://doi.org/10.2307/1938524

Myers JH, Cory JS (2016) Ecology and evolution of pathogens in natural populations of Lepidoptera. Evol Appl 9(1):231–247. https://doi.org/10.1111/eva.12328

Páez D, Fleming-Davies A, Dwyer G (2015) Effects of pathogen exposure on life-history variation in the gypsy moth (Lymantria dispar). J Evol Biol 28(10):1828–1839

Pair SD, Raulston JR, Westbrook JK, Wolf WW, Adams SD (1991) Fall armyworm (Lepidoptera, Noctuidae) outbreak originating in the lower Rio-Grande Valley 1989. Fla Entomol 74(2):200–213

Pitre HN, Hogg DB (1983) Development of the fall armyworm (Lepidoptera, Noctuidae) on cotton, soybean and corn. J Ga Entomol Soc 18(2):182–187

Raymond B, Vanbergen A, Pearce I, Hartley S, Cory J, Hails R (2002) Host plant species can influence the fitness of herbivore pathogens: the winter moth and its nucleopolyhedrovirus. Oecologia 131(4):533–541. https://doi.org/10.1007/s00442-002-0926-4

Richter AR, Fuxa JR, Abdelfattah M (1987) Effect of host plant on the susceptibility of Spodoptera frugiperda (Lepidoptera, Noctuidae) to a nuclear polyhedrosis virus. Environ Entomol 16(4):1004–1006

Shikano I, Shumaker KL, Peiffer M, Felton GW, Hoover K (2017) Plant-mediated effects on an insect-pathogen interaction vary with intraspecific genetic variation in plant defences. Oecologia 183(4):1121–1134

Sparks AN (1979) Review of the biology of the fall armyworm (Lepidoptera, Noctuidae). Fla Entomol 62(2):82–87

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B (Stat Methodol) 64(4):583–639

Szewczyk B, Hoyos-Carvajal L, Paluszek M, Skrzecz I, Lobo de Souza M (2006) Baculoviruses re-emerging biopesticides. Biotechnol Adv 24(2):143–160. https://doi.org/10.1016/j.biotechadv.2005.09.001

Underwood N, Morris W, Gross K, Lockwood JR III (2000) Induced resistance to Mexican bean beetles in soybean: variation among genotypes and lack of correlation with constitutive resistance. Oecologia 122(1):83–89

Underwood N, Rausher M, Cook W (2002) Bioassay versus chemical assay: measuring the impact of induced and constitutive resistance on herbivores in the field. Oecologia 131(2):211–219. https://doi.org/10.1007/s00442-002-0867-y

Vaida F, Blanchard S (2005) Conditional akaike information for mixed-effects models. Biometrika 92(2):351–370

Zens MS, Peart DR (2003) Dealing with death data: individual hazards, mortality and bias. Trends Ecol Evol 18(7):366–373

Sama Shrestha is a graduate student in the Department of Applied Mathematics at University of Colorado Boulder. Her research interests include epidemic modeling and bayesian inference.

Bret D. Elderd is an Associate Professor in the Department of Biological Sciences at Louisiana State University. His research focuses on examining how disease outbreaks, community structure, and environmental variation in uence population dynamics by combining field experiments and theoretical models. He is particularly interested in host-pathogen interactions, variability within and between populations in disease transmission, population viability and rare species management.

Vanja Dukic is a Professor in the Department of Applied Mathematics at University of Colorado Boulder. Her main research interests are in Bayesian modeling, inference, and computational statistics, with applications to a wide variety of fields, ranging from medicine and ecology to risk and insurance. Her work includes modeling of infectious diseases (NPV, smallpox, in uenza, MRSA, and meningitis), IP surveillance, and sequential decision making under uncertainty.

Affiliations

Sama Shrestha[1] · Bret D. Elderd[2] · Vanja Dukic[1]

✉ Sama Shrestha
sama.shrestha@colorado.edu

Bret D. Elderd
elderd@lsu.edu

Vanja Dukic
vanja.dukic@colorado.edu

[1] Department of Applied Mathematics, University of Colorado, Boulder, CO, USA

[2] Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA