Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2024, 12, 24-39
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Modelling the Survival of Western Honey Bee *Apis mellifera* and the African Stingless Bee *Meliponula ferruginea* Using Semiparametric Marginal Proportional Hazards Mixture Cure Model

Patience Isiaho[1,2], Daisy Salifu[2*], Samuel Mwalili[1], Henri E. Z. Tonnang[2]

[1]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[2]Data Management, Modelling, and Geo-Information Unit, International Centre of Insect Physiology and Ecology (icipe), Nairobi, Kenya
Email: *dsalifu@icipe.org

How to cite this paper: Isiaho, P., Salifu, D., Mwalili, S. and Tonnang, H.E.Z. (2024) Modelling the Survival of Western Honey Bee *Apis mellifera* and the African Stingless Bee *Meliponula ferruginea* Using Semiparametric Marginal Proportional Hazards Mixture Cure Model. *Journal of Data Analysis and Information Processing*, **12**, 24-39.
https://doi.org/10.4236/jdaip.2024.121002

Received: November 1, 2023
Accepted: February 1, 2024
Published: February 4, 2024

Abstract

Classical survival analysis assumes all subjects will experience the event of interest, but in some cases, a portion of the population may never encounter the event. These survival methods further assume independent survival times, which is not valid for honey bees, which live in nests. The study introduces a semi-parametric marginal proportional hazards mixture cure (PHMC) model with exchangeable correlation structure, using generalized estimating equations for survival data analysis. The model was tested on clustered right-censored bees survival data with a cured fraction, where two bee species were subjected to different entomopathogens to test the effect of the entomopathogens on the survival of the bee species. The Expectation-Solution algorithm is used to estimate the parameters. The study notes a weak positive association between cure statuses ($\rho_1 = 0.0007$) and survival times for uncured bees ($\rho_2 = 0.0890$), emphasizing their importance. The odds of being uncured for *A. mellifera* is higher than the odds for species *M. ferruginea*. The bee species, *A. mellifera* are more susceptible to entomopathogens icipe 7, icipe 20, and icipe 69. The Cox-Snell residuals show that the proposed semiparametric PH model generally fits the data well as compared to model that assume independent correlation structure. Thus, the semi parametric marginal proportional hazards mixture cure is parsimonious model for correlated bees survival data.

Keywords

Mixture Cure Models, Clustered Survival Data, Correlation Structure, Cox-Snell Residuals, EM Algorithm, Expectation-Solution Algorithm

1. Introduction

The most economically significant pollinators of crop mono-cultures globally continue to be honeybees, especially *Apis mellifera* [1]. According to United Nations *et al.* [2] and Klein *et al.* [3], insects pollinate 75% of the world's crop species and contribute to 35% of food production, which is worth $267 to $657 billion USD yearly, [4]. In addition, honey bees and stingless bees create a variety of hive products, such as honey, wax, cerumen, bee bread, royal jelly, bee venom, and propolis, which are frequently used in cosmetics, pharmaceuticals, and nutrient-rich foods [5].

Honeybee (*Apis mellifera* L.) losses on a large scale have been observed recently all over the globe [6]. Since 2006, reports of Colony Collapse Disorder (CCD), a poorly known syndrome, have been made in the United States. Several possible causes have been claimed for colony losses but there is now a general consensus about the fact that many factors are likely involved. Parasitism by varroa mite (*Varroa destructor*) is considered a major contributor to the collapse of honey bee colonies [7]. Other contributing factors include viral and bacterial infections, poor nutrition, exposure to chemicals used for in-hive pest control, and other agricultural pesticides that bees encountered while foraging [8].

The entomopathogenic fungi *Metarhizium anisopliae* (Metschnikoff) Sorokin and *Beauveria bassiana* (Balsamo) Vuillemin are formulated and used worldwide as biopesticides. These biopesticides are safer alternatives to chemical pest control based on their persistence in the field and environmental compatibility [9]. Despite them being safer it is essential to evaluate the survival of honey bees after being exposed to the entomopathogenic fungi.

Modeling bee survival has been carried out commonly using Kaplan-Meier survival analysis [10], Generalized linear models (GLM) [11] and Cox Proportional Hazard model [12] with the assumption of independent and identically distributed data. However honey bees are social insects meaning that they live together in large, well-organized family groups and survival duration of a bee may depend on the survival duration of its nest-mates resulting into clustered failure time data with potential correlation among survival times within a cluster. Therefore, it is important to take the correlation into account when analyzing clustered failure times.

Most survival techniques also make the assumption that every subject will eventually experience the event of interest. However, there are some circumstances in which a portion of the target population will never encounter the event in question. Survival data typically has heavy right censoring, and a standard

survival analysis model would not always be appropriate. Therefore, the incorporation of a "cured" fraction in a statistical model is necessary. In this paper, we will consider a possible fraction of cured bees. They are cured in the sense that they will not experience death due to exposure to an entomopathogenic fungus and this is modeled via the mixture cure model.

The mixture cure model, first introduced by Boag [13] and Gage [14], is one of the most popular models to estimate the cure rate of treatment and the survival rate of uncured individuals in a study at the same time. The two most common approaches for modeling correlated survival times with a cured fraction are the random effects and marginal models. The random effects models explicitly formulate the underlying dependence structure by frailty and the failure times are assumed to be independent and conditional on the unobservable frailty. The major drawback of the random effects model is that it suffers the need to verify the correlation structure specified by the random effects employed in these models. Because of the fully specified correlation structure, these models are prone to misspecification. To reduce the dependence of a model on the specification of the unobservable correlation structure of clustered data, marginal approaches have been used to handle correlated survival data. The marginal models typically exhibit good robustness to model misspecification because they use a population-average technique to estimate the marginal mean and treat the correlation as nuisance parameters.

Peng *et al.* [15] proposed a semi-parametric marginal proportional hazards mixture cure (PHMC) model to analyze clustered survival data (denoted as the PTY method in this paper). In this model, the correlation within an institution was not explicitly modeled. This marginal model method is robust to correlation misspecification because it does not rely on a particular correlation structure. It is helpful when there is little knowledge of the correlation structure or when the correlation structure is prone to be misspecified. However, using the marginal model may result in an efficiency loss when the correlation is of interest and there is only partial information for the correlation structure accessible.

Niu and Peng [16] proposed a generalized estimating equations (GEE) method based on the Expectation-Solution (ES) algorithm for a semiparametric marginal mixture cure model for clustered failure time data with a possible cure fraction (denoted as NP method in this paper). Their methodology significantly improved the accuracy of the estimation method in Peng *et al.* [15] when the cluster's internal correlation is strong and the cluster size is big. For the regression parameters in the latency, the estimating function in Niu and Peng [16] is biased.

In this study, we will implement a marginal semi-parametric proportional hazards mixture cure model (PHMC) models which was first proposed by Niu and Peng [16] and later extended by Niu *et al.* [17] to the data employing independence and exchangeable correlation structures. This method (denoted as ES method), compared with the existing marginal and random effects approaches con-

sider the correlation structure explicitly in the model as the random effects methods do, but it also enjoys the simplicity of the marginal methods. We use estimating functions for the regression parameters and a semiparametric estimate of the baseline distribution in the latency part of the model.

2. Materials and Methods

2.1. Available Data

This study utilized secondary data from a study that was conducted to assess the impact of entomopathogenic fungi on the Western honey bee (*Apis mellifera* L.) and the African stingless bee (*Meliponula ferruginea* Cockrell), specifically focusing on the nontarget effects [18]. The study included six naturally occurring entomopathogenic fungi from the soil, namely five isolates of *Metarhizium anisopliae* (icipe 7, icipe 20, icipe 62, icipe 69, and icipe 78), and one isolate of *Beauveria bassiana* (icipe 284) evaluated on two bee species. The honey bee species and stingless bee species were studied in cages as they could not be observed individually. Each cage contained approximately 25 - 30 bees. The event of interest in this study was the time-to-death of bees. The experiment was conducted over a period of 10 days, with most bees being right-censored.

2.2. Statistical Modelling

Assume we have K clusters of subjects with n_i individuals in the i^{th} ($i=1,\cdots,K$) cluster. The total number of subjects is $N=\sum_{i=1}^{K} n_i$. Let Y_{ij} represent the subject's cure status for the j^{th} subject in cluster i. $Y_{ij}=1$ if the subject is uncured (susceptible) and 0 otherwise. If the subject is not cured, let $\tilde{T}_{ij}$ be the failure time. The failure time of a cured subject is set at ∞. As a result, $\tilde{T}_{ij}^{*}=Y_{ij}\tilde{T}_{ij}+\left(1-Y_{ij}\right)\times\infty$ is the failure time of a subject. Let C_{ij} be the censoring time for the j^{th} subject in the i^{th} cluster.

For the j^{th} subject in cluster i, let Y_{ij} denote the cure status of the subject with $Y_{ij}=1$ if the subject is uncured (susceptible) and 0 otherwise. Let $\tilde{T}_{ij}$ be the failure time of the subject if the subject is uncured. The failure time of a cured subject is defined at ∞. Therefore, the failure time of a subject is given by $\tilde{T}_{ij}^{*}=Y_{ij}\tilde{T}_{ij}+\left(1-Y_{ij}\right)\infty$. Let C_{ij} denote the censoring time for the j^{th} subject in the i^{th} cluster.

The observed failure time is $\tilde{T}_{ij}=\min\left(T_{ij},C_{ij}\right)$, and its censoring status is denoted as $\delta_{ij}=I\left(T_{ij}\le C_{ij}\right)$, where $I(A)=1$ if A is true and 0 otherwise. Suppose there are two sets of covariates (may share some covariates) X_{ij} and Z_{ij} that may have an impact on the cure probability and the failure time distribution of uncured subjects. Let $S\left(t;x_{ij},z_{ij}\right)$ and $S_u\left(t;x_{ij}\right)$ denote the marginal survival functions of $\tilde{T}*_{ij}$ and $\tilde{T}_{ij}^{*}$, respectively. The marginal PHMC model is defined as:

$$S\left(t;x_{ij},z_{ij}\right)=1-\pi\left(Z_{ij}\right)+\pi\left(Z_{ij}\right)S_u\left(t;x_{ij}\right) \tag{1}$$

where:

$$\pi\left(Z_{ij}\right)=P\left(Y_{ij}=1;Z_{ij}\right)=\frac{\exp\left(\gamma' Z_{ij}\right)}{1+\exp\left(\gamma' Z_{ij}\right)} \tag{2}$$

is the so-called incidence probability model (Cure probability model) and is in a logistic regression form.

$$S_u\left(t;x_{ij}\right)=P\left(T_{ij}>t\mid X_{ij}\right)=\exp\left(-\Lambda_{u0}\left(t\right)\exp\left(\beta' X_{ij}\right)\right) \tag{3}$$

is the latency survival model for uncured subjects (specified by the proportional hazards (PH) model), $\Lambda_{u0}\left(t\right)$ is the cumulative baseline hazard function of $\tilde{T}_{ij}$, and β and γ are two sets of unknown regression parameters with length p_X and p_Z for X_{ij} and Z_{ij}. We consider the semiparametric PHMC models by specifying $\Lambda_{u0}\left(t\right)$ nonparametrically.

2.2.1. Components of the PHMC Model

The marginal PHMC model consists of two parts: the incidence and the latency model. In the context of our study, the incidence model predicts whether a bee will not experience death after exposure to entomopathogens. The latency model predicts the survival time of the bees conditional on the bee being susceptible to death due to the entomopathogens.

2.2.2. Generalized Estimating Equations (GEE)

The Generalized Estimating Equations (GEE) method, grounded in the relative theory, extends the generalized linear model to effectively model correlated observations, such as clustered data. This method, which is based on a quasi-likelihood approach, offers a robust framework for dealing with correlations within clusters, thereby accommodating a variety of correlation structures. These estimating equations, first introduced by Liang and Zeger [19], provide regression estimates for analyzing repeated measures with non-normal response variables. In their pioneering work, they emphasized the flexibility of GEE in handling different types of dependencies in longitudinal data. Later, Niu and Peng [16] proposed a novel estimating equation approach, following the principles laid down by Liang and Zeger [19], to model clustered data with a cured fraction. Their approach, which aligns with the relative theory of GEE, developed estimating functions for a marginal mixture cure model. This methodology involves explicitly specifying the correlation within a cluster and incorporating this correlation directly into the estimating equations, thereby enhancing the model's ability to account for intra-cluster dependencies and improving the accuracy of the estimates.

2.2.3. Estimation of Parameters

Let $\tilde{t}_{ij}$ be the observed value of $\tilde{T}_{ij}$ and $O=\left\{\tilde{t}_{ij},\delta_{ij},x_{ij},z_{ij},i=1,\cdots,K;j=1,\cdots,n_i\right\}$ be the observed data. The PTY method uses the Expectation-Maximization algorithm to estimate parameters while the ES method uses the Expectation-Solution algorithm. The ES estimation method iterates between an expectation step (E-step) and a solution step (S-step) until convergence is achieved. The log-likelihood

$l_c(\gamma,\beta,\alpha:O')$ function based on data O':

$$\begin{aligned} l_c &= \log_{10}\prod_{i=1}^{K}\prod_{j=1}^{n_i}\pi(Z_{ij})^{y_{ij}}\left(1-\pi(Z_{ij})^{1-y_{ij}}\right)\left[f_u(\tilde{t}_{ij};x_{ij})^{\delta_{ij}}S_u(\tilde{t}_{ij};x_{ij})^{1-\delta_{ij}}\right]^{y_{ij}} \\ &= \log_{10}\prod_{i=1}^{K}\prod_{j=1}^{n_i}\pi(Z_{ij})^{y_{ij}}\left(1-\pi(Z_{ij})^{1-y_{ij}}\right) \\ &\quad+\log_{10}\prod_{i=1}^{K}\prod_{j=1}^{n_i}\left(\left(\Lambda_{u0}(\tilde{t}_{ij};\alpha)\exp(\beta' x_{ij})\right)^{\delta_{ij}}\exp\left(-\Lambda_{u0}(\tilde{t}_{ij};\alpha)\exp(\beta' x_{ij})\right)\right)^{y_{ij}} \\ &\quad+\log_{10}\prod_{i=1}^{K}\prod_{j=1}^{n_i}\left(\frac{\lambda_{u0}(\tilde{t}_{ij};\alpha)}{\Lambda_{u0}(\tilde{t}_{ij};\alpha)}\right)^{\delta_{ij}} \end{aligned} \tag{4}$$

where $\lambda_{u0}(t;\alpha)$ and $\Lambda_{u0}(t;\alpha)$ are the corresponding baseline hazard and cumulative baseline hazard functions for $S_{u0}(t)$, and α is a set of unknown parameters in the baseline distribution.

Niu and Peng [16] and Niu *et al.* [17] proposed an approach based on the ES algorithm to estimate $\theta=(\beta,\gamma,\Lambda_{u0})$ in the marginal PHMC model (1). The correlation within clusters is explicitly accounted for and estimated in this approach. Given the current estimates of $\beta,\gamma,\Lambda_{u0}$ at the m^{th} iteration, denoted as θ^m, the E-step calculates the posterior expectation of Y_{ij} as follows:

$$\begin{aligned} g_{ij}^{(m)} &= \mathbf{E}\left(Y_{ij}\mid\theta^{(m)},O'\right) \\ &= \left\{\delta_{ij}+\frac{(1-\delta_{ij})\pi(Z_{ij})\exp\left(-\Lambda_{uo}(t_{ij})\exp(\beta' X_{ij})\right)}{1-\pi(Z_{ij})+\pi(Z_{ij})\exp\left(-\Lambda_{uo}(t_{ij})\exp(\beta' X_{ij})\right)}\right\} \end{aligned} \tag{5}$$

where $O=\{\tilde{t}_{ij},\delta_{ij},x_{ij},z_{ij}\}$ is the observed data.

The S-step updates the estimate of β and γ based on the following generalized estimating equations:

$$U(\gamma)=\sum_{i=1}^{k}U_i(\gamma)=\sum_{i=1}^{k}\left\{\frac{\partial\pi(Z_i)}{\partial\gamma}\right\}\left\{A_i^{1/2}Q_i(\rho_1)A_i^{1/2}\phi_1\right\}^{-1}\left\{g_i^{(m)}-\pi(Z_i)\right\}=0 \tag{6}$$

$$U(\beta)=\sum_{i=1}^{k}U_i(\beta)=\sum_{i=1}^{k}\left\{\frac{\partial\mu(X_i)}{\partial\beta}\right\}\left\{B_i^{1/2}Q_i(\rho_2)B_i^{1/2}\phi_2\right\}^{-1}W_i\left\{k_i-\mu(X_i)\right\}=0 \tag{7}$$

where;

$$g_{ij}^{(m)}=\left(g_{i1}^{(m)},\cdots,g_{in_i}^{(m)}\right)'$$

$$A_i=\operatorname{diag}\left[\pi(Z_{i1})\{1-\pi(Z_{i1})\},\cdots,\pi(Z_{in_i})\{1-\pi(Z_{in_i})\}\right]$$

$$\pi(Z_i)=\left\{\pi(Z_{i1}),\cdots,\pi(Z_{in_i})\right\}'$$

$$\mu(X_i)=\left\{\mu(X_{i1}),\cdots,\mu(X_{in_i})\right\}'$$

$$B_i=\operatorname{diag}\left\{\mu(X_{i1}),\cdots,\mu(X_{in_i})\right\}$$

$$W_i=\operatorname{diag}\left(g_{i1}^{(m)}\Lambda_{uo}(t_{i1}),\cdots,g_{in_i}^{(m)}\Lambda_{uo}(t_{in_i})\right)$$

$$k_i = \left(k_{i1}, \cdots, k_{in_i}\right)'$$

with $\mu X_{ij} = \exp\left(\beta' X_{ij}\right)$, $k_{ij} = \delta_{ij} / \lambda_{uo}\left(t_{ij}\right)$ and $\mathrm{diag}(a)$ is a diagonal matrix with *a* vector *a* as the diagonal elements.

2.2.4. Correlation Structure

We take into account two working correlation structures, *i.e.*, independence and exchangeable (also known as equicorrelated or compound symmetry) [20], for both $Q_i(\rho_1)$ and $Q_i(\rho_2)$ in the estimating Equations (6) and (7).

The independence correlation structure assumes that observations within the same cluster are independent of each other. This implies that the cluster-specific errors are uncorrelated and do not share any common variance. In an exchangeable correlation structure, the correlation between any two observations within the same group or cluster is assumed to be the same. This means that the correlation between any two observations within a group is interchangeable with the correlation between any two other observations within the same group.

The estimated values of ρ_1 and ρ_2 provide good measures of the strength of the correlations between the cure statuses and between the failure times of uncured subjects in a cluster.

2.2.5. Model Evaluation

Model diagnostic techniques for assessing the fit of mixture cure models have received relatively little research. Peng and Taylor [21] developed a number of residual-based model diagnostic tools to assess the fit of the overall model and the latency model for uncured subjects. They can be used to develop cumulative sums of modified martingale residuals to examine the fit of the model, such as the PH assumption.

The three types of model-checking techniques Peng and Taylor [21] developed include; martingale residuals, Cox-Snell residuals and Kolmogorov-type supremum test. In this study, we evaluated the overall fit of the model using the Cox-Snell residuals. According to Peng and Taylor [21] the Cox-Snell residuals for the overall mixture cure model is expressed as:

$$r_i = -\log_{10} S\left(t_i / x_i, z_i\right) = \delta_i - M_i, \ i = 1, \cdots, n \tag{8}$$

where $\left(t_i / x_i, z_i\right)$ is the overall survival function from the mixture cure model and M_i is the Martingale residuals.

Then the Cox-Snell residuals is viewed as a mixed-type distribution with a unit exponential distribution as the continuous component between 0 and $-\log_{10}\left[1 - \pi(z)\right]$ and a probability mass $1 - \pi(z)$ at $-\log_{10}\left[1 - \pi(z)\right]$ as the discrete component. Despite this fact, the standard procedure of comparing the estimated cumulative hazard rate of the $\left(r_i, \delta_i\right)'s$ to the cumulative hazard function of the unit exponential distribution is still valid because the residuals that are equal to $-\log_{10}\left[1 - \pi(z)\right]$ are always censored and the entire residuals can therefore still be regarded as a censored sample from the unit exponential distribution.

The analysis was implemented using R version 4.1.3 [22]. The semi-parametric Cox marginal and frailty models were fitted using *geecure*() function in **geecure** package [23].

3. Results

3.1. Exploratory Data Analysis

Figure 1 shows the Kaplan-Meier survival curve and its pointwise 95% confidence interval is displayed, the curve leveled off at approximately 0.7 for the species *Apis Mellifera* and 0.85 for *Meliponula ferruginea*. This implied that a cure fraction existed in the data and a cure model should be considered. The bees were clustered in cages in the study, which was another significant structure of the data. The shared environment and the treatment administered in one cage may induce a correlation among the failure times of uncured bees in one cage and among the cure statuses. Therefore, it was important a model that takes into consideration of both the cure fraction and the cluster effect.

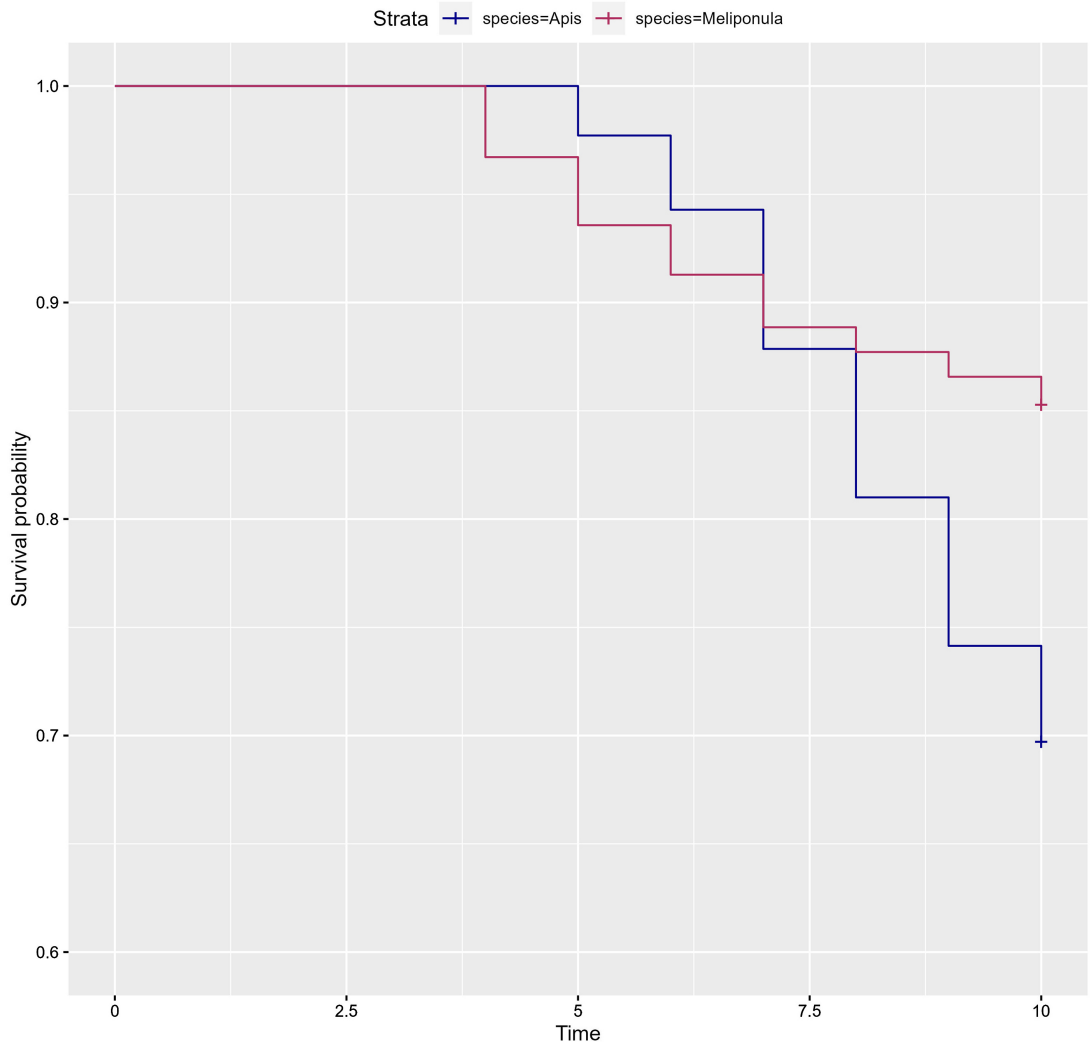

Figure 1. Kaplan-Meier curve for *A. mellifera* and *M. ferruginea* levelling off at different survival probabilities.

3.2. The Semiparametric Marginal PHMC Model

We first fitted a marginal semiparametric PHMC model with an independent correlation structure. The model includes all the covariates (species and treatments) and it has two parts, the logistic model for cure probability and the semiparametric proportional hazards model for the failure time distribution of uncured bees. The results of the fitted model are presented in Table 1. The treatments are very significant in incidence but not in latency. This is because the treatment tends to have a positive effect on the cure probability. It lowers the risk of death to the bees.

Table 1. Parameter and standard errors estimates of the proportional hazard mixture cure model for the bees' survival data under the independence and exchangeable correlation structures.

	Independence (Peng-Taylor-Yu method)			Exchangeable (Expected-Solution method)		
	estimate	SE	p-value	estimate	SE	*p*-value
Cure Probability model:						
Intercept	−1.7030	0.2328	<0.0001	−1.6973	0.2482	<0.0001
Species Meliponula versus Apis)	1.9481	0.1431	<0.0001	1.9487	0.1523	<0.0001
icipe 7	2.0968	0.2748	<0.0001	2.0911	0.2748	<0.0001
icipe 20	2.1214	0.3356	<0.0001	2.1149	0.4632	<0.0001
icipe 62	2.6504	0.2782	<0.0001	2.6507	0.3842	<0.0001
icipe 69	2.2494	0.2895	<0.0001	2.2914	0.3215	<0.0001
icipe 78	2.6825	0.2989	<0.0001	2.6572	0.3012	<0.0001
icipe 284	2.8171	0.2670	0.0972	2.8494	0.2786	0.0724
ρ_1	0				0.0007	
Failure Time Distribution Model: Species (Meliponula versus Apis)						
	0.6484	0.8151	0.4263	0.6232	0.9101	0.7425
icipe 7	0.4074	2.9438	0.8899	0.3890	3.4378	0.8899
icipe 20	0.4962	3.0983	0.8728	0.4672	3.4220	0.8728
icipe 62	0.5034	3.0353	0.8683	0.4003	3.6223	0.8683
icipe 67	0.3175	2.9119	0.9132	0.2974	2.9890	0.9132
icipe 78	0.2107	3.1983	0.9475	0.2657	3.007	0.9475
icipe 284	0.1331	3.5458	0.0430	0.1025	3.5764	0.0414
ρ_2	0				0.0890	

3.2.1. Cure Probability Model

This part of the model assesses the probability of bees being unaffected (or "cured") by the treatments. To interpret these effects of the estimates, for the Cure Probability Model we proceed as for a classical logistic regression model.

The proportion of subjects for which the event does not happen is often called cure rate and is of particular interest especially if there are covariates that are likely to affect it. The cure probability model assesses the probability of bees being unaffected (or "cured") by the treatments as it is our event of interest here. To interpret the estimates of these effects for the Cure Probability Model, we proceed as for a classical logistic regression model.

The cure rate can be calculated based on the results from the Cure probability model part. For example, the cure rate for the *M. ferruginea* is 56.10%, which is calculated by:

$$\text{Cure Rate}_{\text{Meliponula}} = \frac{e^{-1.7030-0.9481}}{1+e^{-1.7030-0.9481}} = 56.10\%$$

While the cure rate for the A mellifera is:

$$\text{Cure Rate}_{\text{Apis}} = \frac{e^{-1.7030}}{1+e^{-1.7030}} = 15.41\%$$

Table 2 shows the cure rates for bees under the different treatments (entomopathogens) under the Independence and Exchangeable correlation structure.

The negative intercept in this model suggests that, at the baseline (with no treatment or for the control group), the log-odds of survival are low, meaning that the probability of survival is less than 0.5. The positive and significant coefficients for species indicate that *M. ferruginea* species had a higher likelihood of being "cured" (unaffected) compared to the *A. mellifera* during the study period under similar treatment conditions. A positive coefficient suggests a higher log-odds of being cured (unaffected) compared to the baseline (no treatment or a control treatment), while a negative coefficient suggests a lower log-odds of being cured. The positive coefficients for icipe treatments (7, 20, 62, 69, 78) suggest that these treatments are less lethal to bees than the baseline treatment (no treatment).

Table 2. Cure rates for the treatments under the independence and exchangeable correlation structure.

	Independence (PTY) estimate	Exchangeable (ES) estimate
icipe 7	59.72%	59.72%
icipe 20	60.31%	60.29%
icipe 62	72.06%	72.18%
icipe 69	63.33%	64.43%
icipe 78	72.70%	72.31%
icipe 284	75.29%	75.99%

The cure rates vary across treatments, with icipe 284, 78, and 62 showing higher survival rates compared to the other treatments. The similarity of the cure rates under both the Independence and Exchangeable models suggests robustness in the results across different correlation structures. Higher cure rates indicate a lower level of lethality associated with the respective treatments. Thus, treatments with higher cure rates like icipe 284, 78, and 62 are less lethal (occurrence of death) to bees in this study.

3.2.2. Failure Time Distribution Model

This part estimates the hazard (or risk) of death for those bees that are not cured that is, bees susceptible or affected by the treatments. For the Failure Time Distribution Model, a Cox PH model is assumed for this part. The coefficients for type of species are not significant (high p-values), suggesting no strong evidence of different survival times between the species. None of the coefficients of The icipe treatments (7, 20, 62, 67, 78, 284) are statistically significant at 0.05 level of significance ($p > 0.05$), indicating no clear evidence that these treatments impact the survival time differently. icipe 284 has a significant p-value in both methods, but given the small coefficient, the effect might be minimal. The correlation coefficient ρ_2 in the exchangeable model is small but significant, indicating some level of correlation in the failure time component of the model.

The marginal PHMC models in the package reduce to the marginal models considered in Peng *et al.* [15], Peng and Taylor [24] (PTY method) when the independence working correlation structure is used in the function. The above conclusions are made based on an analysis that ignores the effect of clusters. That is, all failure times are treated as if they were independent. Ignoring the correlation may lead to incorrect estimates. Therefore it is important to examine the validity of the conclusions above after taking the potential cluster effect into account. We therefore also fitted, the proposed marginal semiparametric PHMC model with an exchangeable correlation structure side by side with the results from the model without taking the potential correlation into account.

It is easy to see that some standard error estimates change substantially when the correlation is taken into account. The consistency of the estimates from the two methods suggests that the correlation within clusters is not significant enough to cause discrepancies in parameter estimations. This is evident from the estimates of ρ_1 and ρ_2. Both values are quite close to zero. They show a weak positive association between the cure statuses and the failure times to the death of uncured bees.

3.3. The Cox-Snell Residuals Plot

Figure 2 presents the Cox-Snell residuals plot from the semiparametric PH mixture cure model. The computed cumulative hazard function doesn't appear to deviate significantly from the 45˚ line. This suggests that even though the residuals' actual distribution is a mixed-type distribution, the unit exponential distribution nevertheless fits them well.

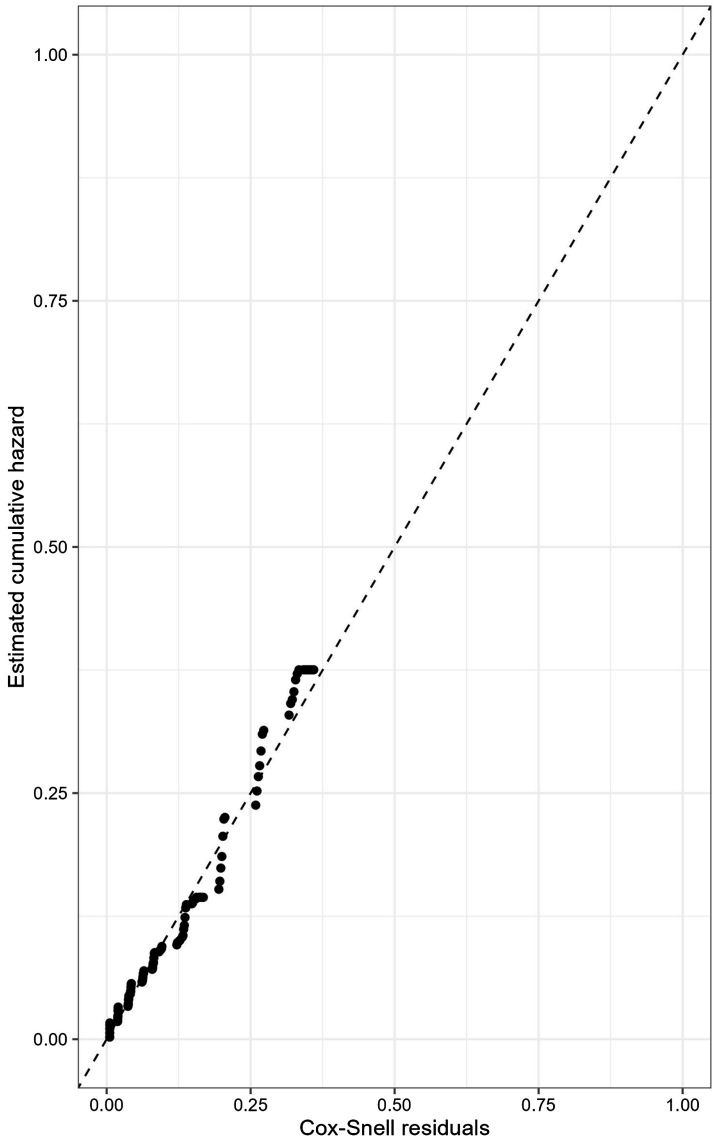

Figure 2. Cox-Snell residuals for the overall marginal PHMC incidence model based on estimated g_{ij}'s (posterior expectation of cure status given observed data) and the weighted Kaplan-Meier estimator.

4. Discussion and Conclusions

This study used a semiparametric marginal proportional hazard mixture model for clustered failure time data with a possible cure fraction proposed by Niu *et al.* [17] (ES method). This model is an extension of Niu and Peng [16] (NP method) model. The model proposes a novel approach on the basis of the generalized estimating equations to incorporate a correlation structure (independent and exchangeable correlation structures) in the marginal model. The marginal PHMC model reduces to the marginal models considered in Niu and Peng [16] when the function uses the independence working correlation structure.

Based on the simulation study done by Niu *et al.* [17] the ES method does not always outperform the PTY [15] and the NP [16] methods. When the correlation

within the cluster is strong and the cluster size is large, the ES method has smaller mean squared errors in $\hat{\beta}$ and $\hat{\gamma}$ than those from the PTY and the NP methods hence performs better [17]. The PTY method appeared to be adequate compared to the ES and the NP methods when the correlation within the cluster was weak. In our study, there was a weak correlation (0.0007 for the cure probability model and 0.0890 for the failure time distribution model) hence using the exchangeable correlation structure produced slightly different estimates compared to the independence correlation structure. However, this suggests that the correlation induced by the clusters among the failure times of uncured bees cannot be ignored.

Peng and Taylor [21] developed a number of residual-based model diagnostic tools to assess the fit of the overall model and the latency model for uncured subjects (Martingale residuals and Cox Snell residuals). These techniques share similar properties as those for the standard survival models and can be used to check the fit of the latency part of a mixture cure model, including functional forms of covariates, identifying outliers, and comparing different mixture cure models. The overall fit of the semiparametric PH mixture cure model with all covariates is examined with the Cox-Snell residuals plotted in **Figure 2**. The residuals show that the proposed semiparametric PH model generally fits the data well.

For clustered survival data with a potential cure fraction, the proposed semiparametric marginal PH mixture model is a useful substitute for the existing marginal models, especially when the correlation structures between cure statuses and between failure times of uncured subjects can be specified up to a few unknown parameters.

Therefore, the proposed semiparametric marginal PH mixture model is a useful alternative to the existing marginal models for clustered data with a possible cure fraction, particularly when the correlation structures among the failure times of uncured subjects among the cure statuses can be specified up to a few unknown parameters. Moreover, unlike a random effects model, a marginal model does not specify explicitly the sources of correlation in the model hence robust to misspecification. A random effects model, as opposed to a marginal model, has advantages when forecasting cluster-specific effects and comprehending the heterogeneity among clusters is of concern.

When the cluster size is large and the correlation within a cluster is strong, as demonstrated by a simulation study done by Niu and Peng [16], the proposed method can significantly enhance estimation efficiency compared to the method used by Peng *et al.* [15]. For clustered survival data with a potential cure fraction, the proposed semiparametric marginal PH mixture model is a useful substitute for the existing marginal models, especially when the correlation structures between cure statuses and between failure times of uncured patients can be specified up to a few unknown parameters.

The computing time is substantial using the **geecure** package. Therefore future studies may be done to reduce the computation time and also explore other

correlation structures such as the first-order autoregressive (AR-1) correlation structure and the unstructured correlation structure.

Acknowledgements

The authors gratefully acknowledge the financial support for this research by the following organizations and agencies: the specific restricted project donor (written out in full) and grant number; the Swedish International Development Cooperation Agency (Sida); the Swiss Agency for Development and Cooperation (SDC); the Australian Centre for International Agricultural Research (ACIAR); the Norwegian Agency for Development Cooperation (Norad); the Federal Democratic Republic of Ethiopia; and the Government of the Republic of Kenya. The first author received a scholarship from the project, Combatting Arthropod Pests for better Health, Food and Climate Resilience (CAP-Africa, Project number: RAF-3058 KEN-18/0005) funded by Norwegian Agency for Development Cooperation (Norad). The views expressed herein do not necessarily reflect the official opinion of the donors.

Availability of Data

The data file and R code used for this study can be found at Bee survival.

Author Contribution

The authors confirm their contribution to the paper as follows: Conceptualization: Daisy Salifu; Methodology: Patience Isiaho, Daisy Salifu, Samuel Mwalili; Software: Patience Isiaho; Validation: Patience Isiaho, Daisy Salifu, Samuel Mwalili; Data Curation: Patience Isiaho; Formal Analysis: Patience Isiaho; Writing - Original Draft: Patience Isiaho; Writing - Review and Editing: Patience Isiaho, Daisy Salifu, Samuel Mwalili, Henri E. Z. Tonnang; Supervision: Daisy Salifu, Samuel Mwalili, Henri E. Z. Tonnang; Funding acquisition: Daisy Salifu, Henri E. Z. Tonnang. All authors reviewed the results and approved the final version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Watanabe, M.E. (1994) Pollination Worries Rise as Honey Bees Decline. *Science*, **265**, 1170-1170. https://doi.org/10.1126/science.265.5176.1170

[2] Food United Nations, Department for Environment, Rural Affairs (DEFRA) Intergovernmental Science-Policy Platform on Biodiversity, and Ecosystem Services (IPBES) (2016) The Assessment Report on Pollinators, Pollination and Food Production: Summary for Policymakers, 8.

[3] Klein, A.-M., Vaissire, B.E., Cane, J.H., Steffan-Dewenter, I., Cunningham, S.A., Kremen, C. and Tscharntke, T. (2007) Importance of Pollinators in Changing Landscapes for World Crops. *Proceedings of the Royal Society B: Biological Sciences*,

274, 303-313. https://doi.org/10.1098/rspb.2006.3721

[4] Porto, R.G., de Almeida, R.F., Cruz-Neto, O., Tabarelli, M., Viana, B.F., Peres, C.A. and Lopes, A.V. (2020) Pollination Ecosystem Services: A Comprehensive Review of Economic Values, Research Funding and Policy Actions. *Food Security*, **12**, 1425-1442. https://doi.org/10.1007/s12571-020-01043-w

[5] Pasupuleti, V.R., Sammugam, L., Ramesh, N. and Gan, S.H. (2017) Honey, Propolis, and Royal Jelly: A Comprehensive Review of Their Biological Actions and Health Benefits. *Oxidative Medicine and Cellular Longevity*, **2017**, Article ID: 1259510. https://doi.org/10.1155/2017/1259510

[6] vanEngelsdorp, D., Evans, J.D., Saegerman, C., Mullin, C., Haubruge, E., Nguyen, B.K., Frazier, M., Frazier, J., Cox-Foster, D., Chen, Y., *et al.* (2009) Colony Collapse Disorder: A Descriptive Study. *PLOS ONE*, **4**, e6481. https://doi.org/10.1371/journal.pone.0006481

[7] Guzmán-Novoa, E., Eccles, L., Calvete, Y., Mcgowan, J., Kelly, P.G. and Correa-Benítez, A. (2010) Varroa Destructor Is the Main Culprit for the Death and Reduced Populations of Overwintered Honey Bee (*Apis mellifera*) Colonies in Ontario, Canada. *Apidologie*, **41**, 443-450. https://doi.org/10.1051/apido/2009076

[8] Van der Sluijs, J.P., Simon-Delso, N., Goulson, D., Maxim, L., Bonmatin, J.-M. and Belzunces, L.P. (2013) Neonicotinoids, Bee Disorders and the Sustainability of Pollinator Services. *Current Opinion in Environmental Sustainability*, **5**, 293-305. https://doi.org/10.1016/j.cosust.2013.05.007

[9] Maina, U., Galadima, I.B., Gambo, F. and Zakaria, D. (2018) A Review on the Use of Entomopathogenic Fungi in the Management of Insect Pests of Field Crops. *Journal of Entomology and Zoology Studies*, **6**, 27-32.

[10] Boff, S., Scheiner, R., Raizer, J. and Lupi, D. (2021) Survival Rate and Changes in Foraging Performances of Solitary Bees Exposed to a Novel Insecticide. *Ecotoxicology and Environmental Safety*, **211**, Article 111869. https://doi.org/10.1016/j.ecoenv.2020.111869

[11] Parish, J.B., Scott, E.S., Correll, R. and Hogendoorn, K. (2019) Survival and Probability of Transmission of Plant Pathogenic Fungi through the Digestive Tract of Honey Bee Workers. *Apidologie*, **50**, 871-880. https://doi.org/10.1007/s13592-019-00697-6

[12] van Dooremalen, C., Gerritsen, L., Cornelissen, B., van der Steen, J.J.M., van Langevelde, F. and Blacquière, T. (2012) Winter Survival of Individual Honey Bees and Honey Bee Colonies Depends on Level of Varroa destructor Infestation. *PLOS ONE*, **7**, e36285. https://doi.org/10.1371/journal.pone.0036285

[13] Boag, J.W. (1949) Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society Series B* (*Methodological*), **11**, 15-44. https://doi.org/10.1111/j.2517-6161.1949.tb00020.x

[14] Berkson, J. and Gage, R.P. (1952) Survival Curve for Cancer Patients Following Treatment. *Journal of the American Statistical Association*, **47**, 501-515. https://doi.org/10.1080/01621459.1952.10501187

[15] Peng, Y., Taylor, J.M.G. and Yu, B. (2007) A Marginal Regression Model for Multivariate Failure Time Data with a Surviving Fraction. *Lifetime Data Analysis*, **13**, 351-369. https://doi.org/10.1007/s10985-007-9042-4

[16] Niu, Y. and Peng, Y. (2013) A Semiparametric Marginal Mixture Cure Model for Clustered Survival Data. *Statistics in Medicine*, **32**, 2364-2373. https://doi.org/10.1002/sim.5687

[17] Niu, Y., Song, L., Liu, Y. and Peng, Y. (2018) Modeling Clustered Long-Term Sur-

vivors Using Marginal Mixture Cure Model. *Biometrical Journal*, **60**, 780-796. https://doi.org/10.1002/bimj.201700114

[18] Omuse, E.R., Niassy, S., Wagacha, J.M., Ong'amo, G.O., Lattorff, H.M.G., Kiatoko, N., Mohamed, S.A., Subramanian, S., Akutse, K.S. and Dubois, T. (2022) Susceptibility of the Western Honey Bee *Apis mellifera* and the African Stingless Bee *Meliponula ferruginea* (Hymenoptera: Apidae) to the Entomopathogenic Fungi *Metarhizium anisopliae* and *Beauveria bassiana*. *Journal of Economic Entomology*, **115**, 46-55. https://doi.org/10.1093/jee/toab211

[19] Liang, K.-Y. and Zeger, S.L. (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22. https://doi.org/10.1093/biomet/73.1.13

[20] Chatterjee, N. and Shih, J. (2001) A Bivariate Cure-Mixture Approach for Modeling Familial Association in Diseases. *Biometrics*, **57**, 779-786. https://doi.org/10.1111/j.0006-341X.2001.00779.x

[21] Peng, Y. and Taylor, J.M.G. (2017) Residual-Based Model Diagnosis Methods for Mixture Cure Models. *Biometrics*, **73**, 495-505. http://www.jstor.org/stable/44695174 https://doi.org/10.1111/biom.12582

[22] R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[23] Niu, Y., Wang, X. and Peng, Y. (2018) geecure: An R-Package for Marginal Proportional Hazards Mixture Cure Models. *Computer Methods and Programs in Biomedicine*, **161**, 115-124. https://www.sciencedirect.com/science/article/pii/S0169260717314542 https://doi.org/10.1016/j.cmpb.2018.04.017

[24] Peng, Y. and Taylor, J.M.G. (2011) Mixture Cure Model with Random Effects for the Analysis of a Multi-Center Tonsil Cancer Study. *Statistics in Medicine*, **30**, 211-223. https://doi.org/10.1002/sim.4098