Likelihood Inference for Flexible Cure Rate Models in the Context of Infectious Diseases with Multiple Exposures

by

ZACHRY J ENGEL

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2020

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Suvra Pal for working with me through graduate school. Dr. Pal taught me what it means to be devoted to your work, the importance of patience and focus in research, and how to guide others. He has been an unwavering guide to me through this process. I would like to thank the members of my dissertation committee, Dr. Shan Sun-Mitchell, Dr. Andrezj Korzeniowski, and Dr. Souvik Roy whose comments through the development of this thesis were invaluable. I wish to thank Dr. Jianzhong Su and Dr. Tuncay Aktosun for helping me receive the Graduate Assistance in Areas of National Need (GAANN) scholarship. This opportunity and the funds it provided were invaluable to me through my graduate school career. I would like to thank all my friends and family who have loved and supported me through my time as a doctoral candidate. I would like to give a special thanks to my parents David and Bernice Engel, who have inspired my entire life to pursue something greater. I would also like to give a special thanks to my friends Ian Lim, Krittamook Kitrungroengkul, Erik Everett, Mark Farinholt, and Mitch Barton for all their support. Finally, and most importantly, I wish to thank my wife Ana, who has been my constant inspiration through all of my collegiate career. Thank you for standing by me through the good times and the bad. Thank you for inspiring me never to give up and always work harder for something better.

August 3, 2020

ABSTRACT

Likelihood Inference for Flexible Cure Rate Models in the Context of Infectious Diseases with Multiple Exposures

Zachry J Engel, Ph.D.

The University of Texas at Arlington, 2020

Supervising Professor: Dr. Suvra Pal

Cure rate models are mostly used to study data arising from cancer clinical trials. Its use in the context of infectious diseases has not been explored well. In 2007, Tournoud and Ecochard first proposed a mechanistic formulation of cure rate model in the context of infectious diseases with multiple exposures to infection. However, they assumed a simple Poisson distribution to capture the unobserved number of pathogens at each exposure time. In this thesis, we propose a new flexible cure rate model to study infectious diseases with discrete multiple exposures to infection. This new model uses the Conway-Maxwell Poisson (COM-Poisson) distribution to model the number of competing pathogens at each moment of exposure. This new formulation takes into account both over-dispersion and under-dispersion with respect to the count on pathogens at each time of exposure and includes the model proposed by Tournoud and Ecochard as a special case. We also propose a new estimation algorithm based on the expectation maximization (EM) algorithm to calculate the maximum likelihood estimates of the model parameters. Infectious diseases data are often right censored, and the EM algorithm can be utilized to efficiently determine the maximum likelihood estimates of the underlying model. We carry out a detailed Monte Carlo simulation study to demonstrate the performance of the proposed estimation algorithm. The flexibility of our proposed model also allows us to carry out a model discrimination, which we do using both likelihood ratio test and information-based criteria. Finally, to illustrate our proposed model, we analyze a recently collected infectious data.

TABLE OF CONTENTS

Appendix

CHAPTER 1

INTRODUCTION

1.1 Cure rate models

The term *cure rate model* refers to models for survival or lifetime data where a portion of the studied population are immune to the event of interest. These cure rate models have become more useful as medicine progresses and finds new treatments that may cause some patients to become non-susceptible to the disease under study. For example, some new treatments for cancers such as leukemia and prostate cancer could cause the patients to be cured of the cancer, also known as going into remission. In the context of cure rate models, those who are no longer susceptible to the disease or the event of interest are known as long-term survivors or immunes. Those who can still be affected by the disease are known as susceptibles. These models represent time-to-event data more realistically than previous methods of survival analysis such as the Cox regression model. In the Cox regression model, we must assume no patients can be cured, and we must only concern ourselves with survival of the patient. Note that not all applications of cure rate models are medical. These models can be extended to fields of study such as engineering, criminology, and sociology. For example, if we wish to apply these cure rate models to the field of criminology, we can look at prisoner recidivism. Recidivism refers to a prisoner returning to prison after being released. If we consider recidivism as our event of interest, then we can consider those who never return to prison as the cured portion of the population of prisoners who have been released. Clearly, knowing what decreases the chance of recidivism is of great importance to the criminal justice system. However, while the implication of cure rate models may be useful to some non-biological fields, the Cox regression model still has its uses in fields such as industrial reliability. Most manufactured goods will fail with time and may need to be replaced. It is therefore advantageous to use Cox regression models or other models that assume no immunes portion to study how long a part will last given certain conditions. While these survival analysis models can be applied to these other fields, most of these models are applied to biomedical research. In particular, the majority of the work in this field has been used to study cancer metastasis and the effectiveness of new treatments to stop the spread of the cancers.

The cause of the event of interest may be due to several factors competing at once. This is known as a competing cause scenario, Cox and Oakes [20]. For example, in the study of cancer, the event of interest may be death which may be caused by the cancer, a stroke, or heart attack brought on by the cancer. In the context of prisoner recidivism, the event of interest is the prisoner coming back to prison which may be caused by a parole violation or a new crime. In both examples, the event of interest is caused by one of the competing causes. A car battery may fail due to cold weather, corrosion, or the chemical reaction in the battery losing power over time. In all these competing cause scenarios, only one of the competing causes was the reason for the event of interest occurring. In fact, only the competing cause which developed first will be the reason for the time-to-event. For example, a normal person may die of a heart attack or a stroke, but not both. We only observe the first to occur. This is a key aspect for competing cause scenarios.

Let M be a random variable denoting the number of competing causes related to the event of interest. Let M have the probability mass function (p.m.f.) $p_m = P[M = m]$ for $m = 0, 1, 2, \cdots$. It is important to note that this probability mass function must include 0 in its support. This will allow us to introduce the cured portion. Given

$M = m$ for $m > 0$, in other words patient is susceptible, let W_i for $i = 1, 2, \cdots, m$ be independent random variables, distributed independently of M, with common distribution function $F(y) = 1 - S(y)$, where $S(y)$ denotes the survival function. The random variable W_i denotes the time taken by the i^{th} competing cause to produce the event of interest, called progression time. As noted previously, not all of these progression times will be observed. In fact, only the first to develop the event of interest will be observed, the rest will be unobserved, or also called latent-variables. To account for those who are not susceptible to the event of interest, the time-to-event, or the lifetime, denoted as Y, will be expressed as follows:

$$Y = \min\{W_0, W_1, \cdots, W_M\}, \tag{1.1}$$

where $P[W_0 = \infty] = 1$. The infinite lifetime W_0 brings in a proportion p_0 of the population who are not susceptible to the event of interest. This proportion is called the *cure rate* and its estimation is of great interest.

1.2 A brief literature review

Due to the wide range of applications, cure rate models and survival analysis have been studied extensively in literature. The first cure rate models were proposed by Boag [12] and Berkson and Gage[11] in which the authors proposed a mixture cure rate model which represented a proportion of the population being cured. Farewell [23] expanded upon this work by considering the mixture model, but employed a logistic regression for the mixture and used a Weibull regression to address the latency. Yakolev et al. [60] created a new formulation of the mixture model to more accurately study cancer metastasis. Yakolev and Tsodikov [59] and later Chen et al. [17] described the promotion time cure rate model by considering a competing cause scenario into their model. Sy and Taylor [54] developed new maximum likelihood techniques for the joint estimation of the incidence and latency promotion time using the competing cause scenario. Yin and Ibrahim [61] proposed a new unified approach to survival analysis with right censored lifetimes. Tucker and Taylor [57] proposed alternatives to the Poisson model, which they saw as having a high rate of error when studying the probability of tumor cure, which they named the *deterministic-stochastic* (DS), *geometric-stochastic* (GS), *Poisson-stochastic*, and*enhanced geometric-stochastic* (GS+) models. Rodrigues et al. [48] were the first to develop a flexible cure rate model where the number of competing causes followed a Conway-Maxwell Poisson (COM-Poisson) distribution. This model can be looked at as a more flexible alternative to the Yin and Ibrahim [61] model. Balakrishnan and Pal [4], Balakrishnan and Pal [5], Balakrishnan and Pal [6], and Balakrishnan and Pal[8] extended the model proposed by Rodrigues et al. [48] by introducing the EM algorithm to find the maximum likelihood estimates of the model and study different lifetime distributions such as the exponential, Weibull, Gamma, and lognormal distributions. This line of research continued with Balakrishnan and Pal [7] in which the authors used the EM algorithm to find the maximum likelihood estimates of a model where the number of competing causes followed the COM-Poisson distribution and the lifetimes were modeled using the generalized Gamma distribution. This model had a high degree in flexibility to model both the lifetime data and the number of competing causes. Other works that expand upon this model include Pal and Balakrishnan [40]. Balakrishnan et al. [10] proposed a semi parametric approach to the COM-Poisson cure rate model by using the Cox proportional hazard model with a Weibull baseline hazard function. Rodrigues et al. [47] developed a flexible cure rate model that included a destructive process, such as chemotherapy for cancer, to the initial risk factors which added more biological and medical context to the models. This model was initially extended by Borges et al. [14] in which the authors used an extension of the generalized power series distribution. Again these destructive models were expanded by Cancho et al. [16] where the authors assumed the number of competing causes followed a negative binomial distribution. Pal and Balakrishnan [39] expanded the destructive model using the negative binomial distribution by incorporating the EM algorithm to find the maximum likelihood estimates. Pal et al. [41] and Majakwara and Pal [37], then incorporated the COM-Poisson distribution into the destructive cure rate model using the EM algorithm. While the COM-Poisson distribution is a natural extension to represent the number of competing causes in most scenarios, there are other works analyzing the effectiveness of other distributions. Gallardo et al. [25] used the Yule-Simon distribution to model the number of competing risks. Gallardo et al. [27] proposed the use of the power series cure rate model. Santos et al. [49] advocated for the use of the Gompertz distribution to model the number of competing causes. Gallardo et al. [26] proposed the use of the polyogarithm distribution for the number of competing causes.

1.3 Lifetime distributions

When studying survival analysis, the researcher must select a distribution to model the time to the event of interest. Throughout history, several distributions have been considered as the true model in a parametric framework. Each model has its advantages and disadvantages when modeling lifetime data. This section will outline some of the more commonly used lifetime distributions.

1.3.1 Exponential distribution

The exponential is the most commonly used lifetime distribution and is the simplest. The probability density function for the exponential distribution is

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0,$$

which gives us the survival function as:

$$S(x;\lambda) = e^{-\lambda x}, \quad x > 0, \lambda > 0.$$

Due to the simplicity of the distribution, the mean and variance are easy to find and are given by

$$E(X) = \frac{1}{\lambda} \text{ and } Var(X) = \frac{1}{\lambda^2}.$$

While the simplicity of the exponential distribution makes it an easy distribution to work with, it does have one distinct disadvantage. The hazard function for the exponential distribution is constant. This makes the exponential distribution less useful than other distributions in practical applications.

1.3.2 Weibull distribution

The Weibull distribution is a continuous distribution named after Waloddi Weibull who first detailed the characteristics of his namesake distribution in 1951. It is one of the most commonly used lifetime models due to its flexibility. The probability density function for the Weibull distribution is

$$f(x;\gamma_1,\gamma_2) = \frac{\gamma_2}{\gamma_1}\left(\frac{x}{\gamma_1}\right)^{\gamma_2-1} e^{(-x/\gamma_1)^{\gamma_2}}, \quad x > 0, \gamma_1 > 0, \gamma_2 > 0.$$

The parameter γ_1 is known as the scale parameter and γ_2 is known as the shape parameter. The survival function for the Weibull distribution is given by

$$S(x;\gamma_1,\gamma_2) = e^{(-x/\gamma_1)^{\gamma_2}}, \quad x > 0, \gamma_1 > 0, \gamma_2 > 0.$$

Due to the greater complexity of the Weibull distribution, the mean and variance are more difficult to calculate and are

$$E(X) = \gamma_1 \Gamma\Big(1 + \frac{1}{\gamma_2}\Big) \text{ and } Var(X) = \gamma_1^2 \Big[\Gamma\Big(1 + \frac{2}{\gamma_2}\Big) - \Big\{\Gamma\Big(1 + \frac{1}{\gamma_2}\Big)\Big\}^2\Big],$$

where $\Gamma(\cdot)$ is the gamma function defined as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx. \tag{1.2}$$

The Weibull distribution is widely used in survival and reliability analyses due to the flexibility of its hazard function. The hazard function of the Weibull distribution can be increasing, decreasing, or constant.

1.3.3 Gamma distribution

The gamma distribution is a two parameter probability distribution commonly used to model lifetime data. An advantage of the gamma distribution is that the exponential, Erlang, and chi-square distributions are special cases of the gamma distribution. The gamma distribution has probability density function

$$f(x; k, \theta) = \frac{1}{\Gamma(\theta)\theta^k} x^{k-1} e^{-x/\theta}, \quad x > 0, k > 0, \theta > 0,$$

where $\Gamma(\cdot)$ is the gamma function as defined in (1.2). Here, k is the shape parameter and θ is the scale parameter. The survival function of the gamma distribution is

$$S(x; k, \theta) = \frac{1}{\Gamma(k)} \gamma\Big(k, \frac{x}{\theta}\Big),$$

where $\gamma(\cdot, \cdot)$ represents the lower incomplete gamma function and is defined as

$$\gamma(z, x) = \int_0^x t^{z-1} e^{-t} dt. \tag{1.3}$$

The mean and variance of the gamma distribution are

$$E(X) = k\theta \text{ and } Var(X) = k\theta^2.$$

1.3.4 Lognormal distribution

Also known as the log-normal or the Galton distribution, is a continuous distribution whose natural logarithm is a normal distribution. This distribution often used to model lifetime data has the probability density function

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\Big(- \frac{(\ln(x)-\mu)^2}{2\sigma^2}\Big), \quad x > 0, \sigma > 0, -\infty < \mu < \infty.$$

The parameters μ and σ are the mean and standard deviation respectively of the normal distribution. The survival function of the lognormal distribution is

$$S(x;\mu,\sigma) = 1 - \Phi\Big(\frac{\ln(x)-\mu}{\sigma}\Big), \quad x > 0,$$

where Φ is the distribution function of the standard normal distribution. The mean and variance of the lognormal distribution are

$$E(X) = e^{\left(\mu+\frac{1}{2}\sigma^2\right)} \text{ and } Var(X) = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}.$$

1.3.5 Generalized Gamma distribution

The generalized gamma distribution is a wider class of distribution which was introduced by Stacy [52] and later expanded on by Prentice [45]. This distribution has an additional shape parameter and the probability density function is given by

$$f(x;a,d,p) = \frac{(p/a^d)x^{d-1}e^{-(x/a)^p}}{\Gamma(d/p)}, \quad x > 0, a > 0, d > 0, p > 0,$$

where $\Gamma(\cdot)$ represents the gamma function as defined in (1.2). The survival function of this distribution is

$$S(x;a,d,p) = 1 - \frac{\gamma(d/p,(x/a)^p)}{\Gamma(d/p)},$$

where $\gamma(\cdot,\cdot)$ represents the lower incomplete gamma function as defined in (1.3). The mean and variance of the generalized gamma distribution are given by

$$E(X) = a\frac{\Gamma((d+1)/p)}{\Gamma(d/p)} \text{ and } Var(X) = a^2\Big[\frac{\Gamma((d+2)/p)}{\Gamma(d/p)} - \Big(\frac{\Gamma((d+2)/p)}{\Gamma(d/p)}\Big)^2\Big].$$

While this distribution is rather complex, it does have some useful aspects. Namely, the exponential, Weibull, and gamma distributions are all special cases of the generalized gamma distribution while the lognormal distribution is a limiting case.

1.4 COM-Poisson cure rate model

The COM-Poisson distribution was introduced by Conway and Maxwell [19] as a solution to handling queuing systems with state-dependent service rates. It is a generalization of the Poisson distribution that adds a dispersion parameter to account for over-dispersed and under-dispersed data relative to the Poisson distribution. We can view the COM-Poisson distribution as a weighted Poisson distribution. This dispersion parameter also allows the geometric and Bernoulli distributions to be special cases of the COM-Poisson distribution. The COM-Poisson distribution was more thoroughly analyzed in the statistical context by Boatwright et al. [13], Shmueli et al. [51], Kokonendji et al. [35], and most recently by Li et al. [36]. Let M follow a COM-Poisson distribution, then:

$$P[M = m; \theta, \nu] = \frac{1}{Z(\theta, \nu)} \frac{\theta^m}{(m!)^\nu}, \quad m = 0, 1, 2, \cdots, \tag{1.4}$$

where $Z(\theta, \nu)$ is known as the normalization constant calculated by:

$$Z(\theta, \nu) = \sum_{j=0}^{\infty} \frac{\theta^j}{(j!)^\nu}. \tag{1.5}$$

From (1.5), we can see the cured fraction, in other words the part of the population that is not susceptible to the event of interest, denoted p_0 is given by:

$$p_0 = P[M = 0; \theta, \nu] = \frac{1}{Z(\theta, \nu)}. \tag{1.6}$$

The COM-Poisson distribution has three distinct special cases. When $\nu = 1$, then $Z(\theta, \nu) = e^\theta$, which results in the Poisson distribution with mean θ. As $\nu \to \infty$,

$Z(\theta,\nu) \rightarrow 1+\theta$, which means the COM-Poisson distribution converges in distribution to the Bernoulli distribution with $P[M = 1; \theta, v] = \frac{\theta}{1+\theta}$. When $\nu = 0$ and $\theta < 1$, $Z(\theta,\nu) = \frac{1}{1-\theta}$, which corresponds to a geometric distribution with parameter $1-\theta$. However, if $\nu = 0$ and $\theta \geq 1$, then $Z(\theta,\nu)$ does not converge. Therefore, the COM-Poisson distribution is undefined in this special case. One advantage of the COM-Poisson distribution is that it allows for both under-dispersion and over-dispersion of count data relative to the Poisson distribution. When $\theta > 1$, the data is under-dispersed relative to the Poisson distribution, whereas if $\theta < 1$, the data is over-dispersed relative to the Poisson distribution. Overdispersion of data occurs when the data exhibits more variation than would be expected in a Poisson distribution with parameter θ. Conversely, underdispersion occurs when the data exhibits less variation than would be expected in a Poisson distribution with parameter θ. This flexibility provides a distinct advantage when modeling count data for competing cause scenarios. With this in mind, let us now examine the long term survival function for the random variable Y in (1.1).

Rodrigues et al. [48] defined the long term survival function of Y as

$$S_{pop}(y) = \frac{Z(\theta S(y),\nu)}{Z(\theta,\nu)}, \tag{1.7}$$

where $Z(\cdot,\cdot)$ is as defined in (1.5). It is important to note that $S_{pop}(y)$ is not a proper survival function since $\lim\limits_{y\rightarrow\infty} S_{pop}(y) = \frac{1}{Z(\theta,\nu)}$. From this long-term survival function, we can find the long-term density function of the random variable Y as

$$f_{pop}(y) = -S'_{pop}(y) = \frac{1}{Z(\theta,\nu)}\frac{f(y)}{S(y)}\sum_{j=1}^{\infty}\frac{j\{\nu S(y)\}^j}{(j!)^\nu}. \tag{1.8}$$

In both (1.7) and (1.8), $S(y)$ and $f(y)$ are the proper survival function and probability density function, respectively, of a lifetime distribution such as those presented in Section 1.3. By taking advantage of the COM-Poisson's relationship to the geometric, Poisson, and Bernoulli distributions, we can use (1.6), (1.7), and(1.8) to find the long-term survival function, long-term probability density function, and cure rate, respectively, for the geometric, Poisson, and Bernoulli models simply by adjusting the dispersion parameter ν. In Table 1.1 we present these functions for the COM-Poisson cure rate model and its three special cases.

Table 1.1. Long-term survival function (S_{pop}), long-term density function (f_{pop}), and cured portion (p_0) for the COM-Poisson cure rate model and its three special cases

Model	$S_{pop}(y)$	$f_{pop}(y)$	p_0
COM-Poisson	$\frac{Z(\theta S(y),\nu)}{Z(\theta,\nu)}$	$\frac{1}{Z(\theta,\nu)} \frac{f(y)}{S(y)} \sum_{j=1}^{\infty} \frac{j\{\nu S(y)\}^j}{(j!)^\nu}$	$\frac{1}{Z(\theta,\nu)}$
Poisson	$e^{-\theta F(y)}$	$\theta f(y) e^{-\theta F(y)}$	$e^{-\theta}$
Bernoulli	$\frac{1+\theta S(y)}{1+\theta}$	$\frac{\theta}{1+\theta} f(y)$	$\frac{1}{1+\theta}$
Geometric	$\frac{1-\theta}{1-\theta S(y)}$	$\frac{(1-\theta)\theta f(y)}{(1-\theta S(y))^2}$	$1-\theta$

1.5 Form of data

For this thesis, we consider the situation where the lifetime in (1.1) is not completely observed and is thus right censored. In a sample size n, let C_i denote the right censoring time and Y_i denote the actual lifetime as described in (1.1) for $i = 1, \cdots, n$. Let $T_i = \min\{Y_i, C_i\}$ denote the observed lifetime for the i^{th} subject. Let δ_i denote a censoring indicator such that $\delta_i = 1$ if $T_i = Y_i$ and $\delta_i = 0$ if $T_i = C_i$. In other words, $\delta_i = 1$ if we observe the true lifetime and $\delta_i = 0$ if the lifetime is right censored. This leaves us with an ordered pair of numbers (T_i, δ_i) for each subject $i = 1, \cdots, n$ to represent their lifetimes.

While most data pertaining to survival analysis and cure rate models are right censored, there are other forms of censoring. Interval censoring is another common form of censoring in which the exact time of event is unknown, but it is known that the event occurred within some interval of time. This happens when the subjects are not under continuous observation, but are rather observed at regular intervals of time. The use of interval censoring in the context of COM-Poisson cure rate models has been investigated by Pal and Balakrishnan [40] and Wiangnak and Pal [58]). Another common form of censoring is left censoring in which case the unit has failed before proper measurements of failure time have begun. Interval and left censoring are more commonly used in fields such as industrial reliability. Another mechanism of censoring worth mentioning is informative and non-informative censoring. This type of censoring is based on whether the lifetime of subjects is dependent or independent of the censoring mechanism. For example, a patient may withdraw from a clinical trial because his/her condition is deteriorating and may need a different treatment. In this case, the patients may expect death to be sooner and as such, the right censoring is informative. On the other hand, if a patient withdraws from a clinical trial because he/she moves to a different place, the right censoring does not provide any information on the patient's lifetime. Hence, in this case, the right censoring is non-informative. Examples of all three of these censoring mechanisms can be found in works such as Kim and Jhun [33] in which the authors considered interval censoring for cure rate models, Hough et al. [30] in which the authors used left censoring to apply survival analysis to food shelf life, and Campigotto and Weller [15] in which the authors analyzed the impact of informative censoring on the Kaplan-Meier estimate of survival.

1.6 Likelihood inference

The most important aspect of a parametric statistical model is the estimation of the unknown parameters of the underlying statistical model. In a parametric framework, we assume the data follows a known distribution and attempt to find the parameters of that distribution that best fit the data at hand. The likelihood function is employed to find these unknown parameters. We use the likelihood function because the likelihood principle states that all the information of the unknown parameters of an underlying function is contained when the data is observed. Maximum likelihood estimation is the most common way to find these unknown parameters. The maximum likelihood technique finds the unknown parameters of the parametric distribution by maximizing the likelihood function. This method provides a unified approach to find the unknown parameters. However, a closed form of this maximization may not be available or the maximum likelihood estimates (MLE) may not exist at all. In such a case, we may employ a numerical technique to find the MLE such as the Newton-Raphson.

An issue that arises with our data is the constant presence of censoring which results in missing data. As stated previously, our data is right censored, and we must utilize an approach that takes into account this missing data. The technique we will employ to address this issue is the expectation maximization (EM) algorithm (Dempster et al. [22]). The EM algorithm is an iterative algorithm that handles missing data quite well while finding the unknown parameters.

1.6.1 EM algorithm

We will use the EM algorithm to carry out the maximum likelihood estimation of model parameters. First, let F_{pop} and F_1 denote the cumulative distribution function (c.d.f.) of the entire population and susceptible population respectively. Furthermore, let S_{pop} and S_1 denote the survival function of the entire population and susceptible population respectively. Let J denote the latent cured status variable which takes on the value of 0 if the subject is immune and 1 if the subject is susceptible. As a consequence, we have $P[J = 0] = p_0$ $P[J = 1] = 1 - p_0$. Then, we have

$$F_{pop}(y) = (1 - p_0)F_1(y),$$

and

$$S_{pop}(y) = 1 - F_{pop}(y) = p_0 + (1 - p_0)S_1(y).$$

Using the form of $S_{pop}(y)$ as in (1.7), we can get an expression for S_1 as

$$S_1(y) = \frac{S_{pop}(y) - p_0}{1 - p_0}.$$

Let δ denote the censoring indicator such that $\delta = 1$ when the actual lifetime is observed and $\delta = 0$ when the lifetime is right censored. Let us define two sets I_1 and I_0 as $I_1 = \{i : \delta_i = 1\}$ and $I_0 = \{i : \delta_i = 0\}$. Note that for the set I_0, the value of J is unknown, and this introduces missing data. Let the complete data be denoted by $(y_i, \delta_i, \mathbf{x}_i, J_i)$ for $i = 1, 2, \ldots, n$, which includes both observed and unobserved J_i 's with $\mathbf{x}_i$ representing a vector of covariates. Let $\boldsymbol{\theta}$ denote the parameter vector of our model. Then, the complete data likelihood function is given by

$$L_c(\boldsymbol{\theta}) = \prod_{I_1}\{f_{pop}(y_i)\}\prod_{I_0}\{p_{0i}\}^{1-J_i}\{(1 - p_{0i})S_1(y_i)\}^{J_i}.$$

Given the above likelihood function, we can write the log-likelihood function as:

$$\begin{aligned} l_c(\boldsymbol{\theta}) = &\sum_{I_1}\log\{f_{pop}(y_i)\} + \sum_{I_0}(1 - J_i)\log\{p_{0i}\} \\ &+ \sum_{I_0} J_i \log\{(1 - p_{0i})S_1(y_i)\}. \end{aligned} \tag{1.9}$$

Now that we have expressed our log-likelihood function, we can begin discussion of the EM algorithm; see McLachlan and Krishnan [38]. In the expectation step (E-step), we calculate the expectation of the complete log-likelihood function with respect to the distribution of the unobserved J_i's, given the model parameters and the observed data. We should note here that J_i's are Bernoulli random variables and are linear in the complete data log-likelihood function. As such, at the r^{th} iteration step, we simply need to calculate $\pi_i^{(r)} = E(J_i|\boldsymbol{\theta}^{(r)}, \text{data})$, for $i \in I_0$, where $\boldsymbol{\theta}^{(r)}$ denotes the current parameter value at the r^{th} iteration of the EM algorithm. Hence, for the i^{th} censored observation, we can calculate $\pi_i^{(r)}$ as

$$\begin{aligned}
\pi_i^{(r)} &= P[J_i = 1|Y_i > y_i; \boldsymbol{\theta}^{(r)}] \\
&= \frac{P[Y_i > y_i|J_i = 1]P[J_i = 1]}{P[Y_i > y_i]}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}} \\
&= \frac{(1 - p_{0i})S_1(y_i)}{S_{pop}(y_i)}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}} \\
&= \frac{S_{pop}(y_i) - p_{0i}}{S_{pop}(y_i)}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}} \\
&= 1 - \frac{p_{0i}}{S_{pop}(y_i)}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}}.
\end{aligned}$$

Therefore, in the E-step, we only replace J_i in (1.9) with $\pi_i^{(r)}$ if the i^{th} observation is censored. As done in McLachlan and Krishnan [38], we will denote the conditional expectation of the complete data log-likelihood function at the r-th iteration as $Q(\boldsymbol{\theta}, \boldsymbol{\pi}^{(r)})$, where $\boldsymbol{\pi}^{(r)}$ is the vector of $\pi_i^{(r)}$ values.

The next step in the EM algorithm is the maximization step (M-step). In this step, we maximize $Q(\boldsymbol{\theta}, \boldsymbol{\pi}^{(r)})$ with respect to $\boldsymbol{\theta}$ over the parameter space $\boldsymbol{\Theta}$. In other words, we choose $\boldsymbol{\theta}^{(r+1)}$ to be a value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that

$$\boldsymbol{\theta}^{(r+1)} = \arg \max_{\theta \in \Theta} Q(\boldsymbol{\theta}, \boldsymbol{\pi}^{(r)}).$$

We continue to repeat this algorithmic process until some convergence criterion is satisfied, for example, $|\frac{\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}}{\boldsymbol{\theta}^{(r)}}| < \epsilon$, where ϵ is some pre-defined tolerance such as $\epsilon = 0.001$.

1.7 Thesis structure

The rest of the thesis will be organized as follows. In Chapter 2, we will explore the use of cure rate models in the context of infectious diseases. We will also introduce the concept of multiple discrete exposures and propose a new cure rate model utilizing the COM-Poisson distribution. In Chapter 3, we will describe in detail the process for generating data from our proposed model for the purpose of a Monte Carlo simulation study. In chapter 4, we will present the results of our simulation study to show the performance of the developed EM algorithm in retrieving the true parameter values of our proposed model. We will also develop model discrimination and model selection procedures through the use of likelihood ratio test and information-based criteria. In chapter 5, we will use our proposed model to analyze a real data pertaining to patients exposed to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Finally, in Chapter 6 we will present some closing remarks and discuss future research opportunities in this direction.

CHAPTER 2

CURE RATE MODELS FOR INFECTIOUS DISEASES

2.1 Previous work

As mentioned in the introduction, cure rate models and survival analysis as a whole can be applied to multiple fields of study. However, the vast majority of the work in cure rate models has been devoted to the study of cancer, whether that is studying possible causes of the disease spreading, evaluating new treatments to slow or stop the progression, and the effectiveness of drugs for treatment of symptoms and causes. However, very little work has been done using cure rate models to study infectious diseases. While the framework of survival analysis is well suited for the study of the progression and spread of an infectious disease, such as human immunodeficiency virus (HIV), the research for infectious diseases is not a prevalent as studies involving cancer. Some of the first work done using survival analysis for infectious diseases was done by Panjer [43] where the authors studied the survival times of patients with HIV which eventually progresses into acquired immunodeficiency syndrome (AIDS). This line of work continued with Green et al. [28] in which the authors studied the survival times of hemophilia-associated AIDS. For historical context, AIDS was a main focus of early research of infectious diseases using survival analysis since the "AIDS Epidemic" began in the 1980's, around the same time survival analysis was being more extensively developed. Struthers and Farwell [53] were the first to apply mixture models to study the chance of contracting HIV or AIDS after an exposure. The use of survival analysis to study the survival rates of patients with HIV or AIDS, which are directly linked to each other, continued with works such as Chequer et al. [18], Jewell and Kalbfleisch [32], and Faucett et al. [24]. While there has certainly been a large effort into using survival analysis to study HIV and AIDS, it is by no means the only infectious disease studied using survival analysis. Hagan et al. [29] used survival analysis to study the lifetimes of patients with Hepatitis C, a dangerous infection of the liver. Ravi et al. [46] studied the survival of cats with feline immunodeficiency virus (FIV) in Canada. All these works have contributed to the advancement of the field of survival analysis to study infectious diseases.

2.2 Cure rate models with multiple exposures

There is an issue with the previous work in survival analysis in the context of infectious diseases; previous works have assumed there was a single point of exposure to the disease, which is not always the case, especially in the context of an infectious disease. For example, HIV and AIDS are spread by the exchange of bodily fluids. While it is possible to contract the infection through a single exposure, those who are sexually active may come into contact with several partners who have the infection, making the likelihood of contracting the disease far more likely. Another example pertains to the 2020 pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), also known as COVID-19 or simply coronavirus. This pathogen is highly contagious and causes flu like symptoms and respiratory distress. Since this disease is so widespread and does not require physical contact between people to spread, many people are exposed several times before contracting the disease. While several researchers were aware of this issue, the first proposed model to study the spread of an infectious disease from multiple exposures was created by Tournoud and Ecochard [55]. This new model allowed the authors to study the survival rates of patients who were exposed to the disease at multiple points in time. This new two component mixture model created a flexible framework that allowed the authors to study models in which there was a single exposure, or multiple. However, the authors did not assume the distribution representing the number of pathogens that entered the body at each moment of infection were identically distributed. Rather, they created a model in which each exposure time is allowed to vary depending on the moment of infection and the patient being infected. For example, one of situations examined by the authors involved nosocomial, or hospital acquired, urinary tract infections (UTI) caused by catheters. The authors considered the initial insertion of the catheter as one level of exposure, and each day after a different level of exposure. In other words, the initial insertion of the catheter most likely exposes the patient to far more bacteria that can cause a UTI than the patient would by just having the catheter in on a day-to-day basis. This flexibility is a great advantage when modeling different infectious diseases. The authors considered the event of interest to be the pathogen promotion time, which refers to the amount of time that passes between a moment of exposure to the pathogen and the first biological signs of infection caused by the pathogen. In everyday life, we are exposed to thousands of bacteria and viruses which try to grow in our body. However, our immune system is able to fight off most infections and thus will show no signs of the possibly dangerous infection. This allows the authors to study an important event of interest and naturally introduce an immune portion of the population. We will now describe the proposed model used by the authors.

Let n represent the number of subjects in the study. Let $t = \{t_0, ..., t_k, ...t_T\}$ denote the multiple and successive moments of infections with t_0 denoting the initial moment of infection. At each moment of infection, let M_{i,t_k}, for $1 \leq i \leq n$, $t_0 \leq t_k \leq t_T$, denote the number of pathogens infecting the i-th subject at time t_k. Note that M_{i,t_k} are unobservable variables, and we assume it to follow a discrete distribution with mass function $p_{M_{i,t_k}}$. As mentioned previously, we do not assume these distributions representing count data to be identically distributed, rather, at each exposure time, we assume a suitable parameter of the distribution to depend on the exposure time. Let Z_{i,j,t_k}, for $1 \leq i \leq n$, $t_0 \leq t_k \leq t_T$ and $0 \leq j \leq M_{i,t_k}$, be the j-th pathogen promotion time of the i-th subject at time t_k, in other words, the time taken by the j-th pathogen to produce the event, which in the context of infectious diseases may represent the first biological sign of infection. For a given M_{i,t_k}, we assume Z_{i,j,t_k} to be distributed with distribution function $F(z) = F(z|\boldsymbol{\gamma}) = 1 - S(z|\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is the associated vector of parameters and $S(\cdot)$ is the corresponding survival function. Since this model follows data with multiple exposures, we need to re-parameterize the time-scale of infection on the original scale. Let s_k denote the time between the initial exposure time t_0 and the k^{th} exposure time t_k. Let $y = z + s_k$, then we can re-parameterize the cdf and survival function as follows:

$$F_{t_k}(y|\boldsymbol{\gamma}) = \begin{cases} F(y - s_k), & y - s_k > 0 \\ 0, & y - s_k \leq 0 \end{cases} \tag{2.1}$$

and

$$S_{t_k}(y|\boldsymbol{\gamma}) = \begin{cases} S(y - s_k), & y - s_k > 0 \\ 1, & y - s_k \leq 0. \end{cases} \tag{2.2}$$

Given that patients are exposed at different discrete time points and at each exposure time there are several competing pathogens, the time-to-event, which can be the first biological sign of infection or time to recovery from an infectious disease, can be defined as

$$Y_i = \min\{Z_{i,j,t_k}, 0 \leq j \leq M_{i,t_k}, t_0 \leq t_k \leq t_T\}, \quad i = 1, 2, \cdots, n, \tag{2.3}$$

where $Z_{i,0,t_k}$ is such that

$$P(Z_{i,0,t_k} = \infty) = 1, \quad t_0 \leq t_k \leq t_T.$$

A patient is termed “immune” of an infection if there are no competing pathogens at each exposure time point, and its probability, termed as cure rate, is defined as

$$p_{0i} = P(M_{i,t_k} = 0 \;\; \forall k \in \{0, 1, 2..., T\}), \quad i = 1, 2, \cdots, n.$$

In their paper, Tournoud and Ecochard [55] assumed $p_{M_{i,t_k}}$ to follow a Poisson distribution with mean θ_{t_k}. While the Poisson distribution is commonly used and certainly has its advantages to model count data, the mean and the variance of the Poisson distribution are identical, meaning the Poisson distribution is not suited to handle over- and under- dispersed data. Furthermore, the authors provided no justification for the suitability of the use of the Poisson distribution to model the number of pathogens at each moment of infection. To address these issues and to expand upon their work, Tournoud and Ecochard [56] studied their previous model, but assumed the number of pathogens the patient is exposed to at each moment of infection followed different distributions. In this paper, the authors considered the same framework for their model as described above, but introduced the Bernoulli, negative binomial, and Compound Poisson distribution to model the number of pathogens entering the body at time t_k. In other words, these new distributions were used to model $p_{M_{i,t_k}}$. These new distributions gave their model a new outlook on the count data, which in turn helped more accurately predict the immune portion of the sample. The most intriguing of the new distributions considered by the authors was the Compound Poisson distribution. The Compound Poisson distribution with parameters $\theta \geq 0$, $\gamma_1 > 0$, and $\gamma_2 > 0$, is defined as

$$M = \begin{cases} X_1 + X_2 + \cdots + X_N & \text{if } N > 0 \\ 0 & \text{if } N = 0, \end{cases} \tag{2.4}$$

where N follows a Poisson distribution with mean θ and $X_1, X_2, \cdots$ are independent and identically distributed random variables from the Gamma distribution with parameters γ_1 and γ_2. While this model does allow for far more flexibility with regard to the count data, the biological interpretation of this model is not clear to see. The issue with using the Compound Poisson distribution to model biological count data is the interpretation of a distribution in which a discrete random variable is attained using the sum of continuous random variables. While this distribution certainly has its uses for other branches of mathematics, such as stochastics, its uses in a biological context are limited.

2.3 COM-Poisson cure rate model with multiple exposures

To address the issue with Compound Poisson distribution and add more flexibility to the model, we propose the use of the COM-Poisson distribution to model the count data of the number of pathogens at each exposure time. Unlike the Compound Poisson distribution, the COM-Poisson is a true discrete distribution, which eliminates the issue of interpretation caused by the Compound Poisson distribution. Furthermore, as we stated in Section 1.4, the Poisson, geometric, and Bernoulli distributions are all special cases of the COM-Poisson distribution, obtained by adjusting the dispersion parameter ν. Also, by adjusting the dispersion parameter further, we can account for data that is over-dispersed or under-dispersed relative to the Poisson distribution. All of these factors lead to a model with more flexibility than those previously studied. The COM-Poisson model will allow us to more accurately model data and allow for a more clear perspective of the true cure rates in a group of subjects.

In this research, we will use the same framework for studying multiple exposures as described in the previous sub-section, but we will assume the number of competing pathogens, $p_{M_{i,t_k}}$, to follow a COM-Poisson distribution with parameters θ_{t_k} and ν at each exposure time t_k, for $k = 0, 1, 2, \ldots, T$. Note that the parameter θ_{t_k} represents the infection intensity at each exposure time t_k and hence carries biological interpretation. Furthermore, we will assume the dispersion parameter ν to be identical for all subjects in the study. In a practical scenario, the infection intensity at a given exposure time will differ across patients. To capture this heterogeneity in patient population, we propose to link θ_{t_k} to a set of covariates $\boldsymbol{x}_{t_k}$ at each exposure time t_k, for $k = 0, 1, 2, \ldots, T$, using the log-linear link function $\theta_{t_k} = \exp(\boldsymbol{x}'_{t_k}\boldsymbol{\beta}_{t_k})$, where $\boldsymbol{\beta}_{t_k}$ is the corresponding vector of regression coefficients. Note that the log-linear link function is not a viable option when considering the geometric distribution. As stated previously in, for the COM-Poisson distribution to converge to the geometric distribution, we need $\nu = 0$ and $0 < \theta_{t_k} < 1$, which is not guaranteed while using the log-linear link function. As such, we will not study the geometric distribution in detail as a special case of the COM-Poisson distribution.

Now, let us consider two exposure times, t_0 and t_1, and derive the survival function of the random variable Y in (2.3), known as the population survival function or the long-term survival function.

Theorem 2.3.1. *Given two discrete exposure times t_0 and t_1, let M_{t_0} and M_{t_1} denote the number of pathogens at times t_0 and t_1. Let M_{t_0} and M_{t_1} both follow COM-Poisson distribution with parameters (θ_{t_0}, ν) and (θ_{t_1}, ν), respectively. Furthermore, let $S_{t_0}(y)$ and $S_{t_1}(y)$ denote the pathogen promotion time distribution at exposure times t_0 and t_1,* *respectively, as described in* (2.2)*. Then, the overall survival function of the variable* Y *in* (2.3)*, denoted* $S_{pop}(y)$*, is given by*

$$S_{pop}(y) = P[Y > y] = \frac{Z(\theta_{t_0} S_{t_0}(y), \nu)}{Z(\theta_{t_0}, \nu)} \frac{Z(\theta_{t_1} S_{t_1}(y), \nu)}{Z(\theta_{t_0}, \nu)}. \tag{2.5}$$

A proof of Theorem 2.3.1 is provided in the Appendix.

Corollary 2.3.1.1. *If we generalize the survival function in Theorem 2.3.1 with multiple exposure times* $t = \{t_0, \ldots, t_k, \ldots t_T\}$*, then, we have*

$$S_{pop}(y) = \prod_{k=0}^{T} \frac{Z(\theta_{t_k} S_{t_k}(y), \nu)}{Z(\theta_{t_k}, \nu)}. \tag{2.6}$$

Note that in (2.6), if we just consider one exposure, then, $S_{pop}(y) = \frac{Z(\theta_{t_0} S_{t_0}(y), \nu)}{Z(\theta_{t_0}, \nu)}$, which reduces to the model proposed by Rodrigues et al. [48] and Balakrishnan and Pal [3]. The density function corresponding to (2.6), known as the long-term density function, is given by

$$f_{pop}(y) = -S'_{pop}(y) = \sum_{k=0}^{T} \left[\frac{1}{Z(\theta_{t_k}, \nu)} \frac{f_{t_k}(y)}{S_{t_k}(y)} \sum_{j=1}^{\infty} \frac{j\{\theta_{t_k} S_{t_k}(y)\}^j}{(j!)^v} \prod_{\substack{i \neq k \\ i=0}}^{T} \frac{Z(\theta_{t_i} S_{t_i}(y), \nu)}{Z(\theta_{t_i}, \nu)} \right], \tag{2.7}$$

where $f_{t_k}(y)$ is the density function associated with $S_{t_k}(y)$. Hence, the cured fraction is given by

$$p_0 = S_{pop}(\infty) = P[M_{t_k} = 0 \ \ \forall k \in \{0, 1, 2..., T\}] = \prod_{k=0}^{T} \frac{1}{Z(\theta_{t_k}, \nu)}. \tag{2.8}$$

With this information and the special relationship the COM-Poisson distribution has with the geometric, Poisson, and Bernoulli distributions, we can adjust the dispersion parameter ν in (2.6), (2.7), and (2.8) to find the long-term survival function, long-term density function, and cured proportion, respectively, for the geometric, Poisson, Bernoulli cure rate models with multiple exposures. In Table 2.1, we present the long-term survival function and the cured proportion for the COM-Poisson cure rate model and its special cases with discrete multiple exposures. In Table 2.2, we present the corresponding long-term density functions.

Table 2.1. Expressions of long-term survival function and cured proportion for the COM-Poisson cure rate model and its special cases with discrete multiple exposures

Model	$S_{pop}(y)$	p_0
COM-Poisson	$\prod_{k=0}^{T} \frac{Z(\theta_{t_k} S_{t_k}(y),\nu)}{Z(\theta_{t_k},\nu)}$	$\prod_{k=0}^{T} \frac{1}{Z(\theta_{t_k},\nu)}$
Poisson	$\exp\Big[\sum_{k=0}^{T}\{-\theta_{t_k} F_{t_k}(y)\}\Big]$	$\exp\Big[\sum_{k=0}^{T}\{-\theta_{t_k}\}\Big]$
Bernoulli	$\prod_{k=0}^{T}\left\{\frac{1+\theta_{t_k} S_{t_k}(y)}{1+\theta_{t_k}}\right\}$	$\prod_{k=0}^{T}\left\{\frac{1}{1+\theta_{t_k}}\right\}$
Geometric	$\prod_{k=0}^{T} \frac{1-\theta_{t_k}}{1-\theta_{t_k} S_{t_k}(y)}$	$\prod_{k=0}^{T}(1-\theta_{t_k})$

Table 2.2. Expressions of long-term density function for the COM-Poisson cure rate model and its special cases with discrete multiple exposures

Model	$f_{pop}(y)$
COM-Poisson	$\sum_{k=0}^{T}\left[\frac{1}{Z(\theta_{t_k},\nu)} \frac{f_{t_k}(y)}{S_{t_k}(y)} \sum_{j=1}^{\infty} \frac{j\{\theta_{t_k} S_{t_k}(y)\}^j}{(j!)^v} \prod_{\substack{l\neq k \\ l=0}}^{T} \frac{Z(\theta_{t_l} S_{t_l}(y),\nu)}{Z(\theta_{t_l},\nu)}\right]$
Poisson	$\sum_{k=0}^{T}\{\theta_{t_k} f_{t_k}(y)\}\exp\Big[\sum_{l=0}^{T}\{-\theta_{t_l} F_{t_l}(y)\Big]$
Bernoulli	$\sum_{k=0}^{T}\left[\left\{\frac{\theta_{t_k}}{1+\theta_{t_k}} f_{t_k}(y)\right\} \prod_{\substack{l\neq k \\ l=0}}^{T}\left\{\frac{1+\theta_{t_l} S_{t_l}(y)}{1+\theta_{t_l}}\right\}\right]$
Geometric	$\sum_{k=0}^{T}\left[\frac{(1-\theta_{t_k})\theta_{t_k} f_{t_k}(y)}{(1-\theta_{t_k} S_{t_k}(y))^2} \prod_{\substack{l\neq k \\ l=0}}^{T} \frac{1-\theta_{t_l}}{1-\theta_{t_l} S_{t_l}(y)}\right]$

The above expressions will be required to construct the complete data likelihood function, which is essential to find the MLEs of the model parameters. By taking advantage of the COM-Poisson's relationship to the other three distributions, we would only need to model the data using the COM-Poisson distribution and then adjust ν according to which distribution we are trying to fit to the data.

2.4 Maximum likelihood estimation

We will now discuss our proposed method for the maximum likelihood estimation of the parameters of our model. To do so, we will employ the EM algorithm. We have previously discussed the EM algorithm in Section 1.6.1 of Chapter 1. For the sake of brevity, we will not repeat the information again. We will now present the expressions for the Q-function, $Q(\boldsymbol{\theta}, \boldsymbol{\pi}^{(r)})$, which was described in Section 1.6.1. To express the Q-function fully, we must first discuss the distribution of the pathogen promotion times, in other words the amount of time it takes for the pathogen to produce the event, for our model. In our model, we propose the use of the Weibull distribution to model the pathogen promotion time at each exposure time. Let γ_1 denote the shape parameter of a Weibull distribution. If $\gamma_1 < 1$, then the failure rate will decrease over time. If $\gamma_1 = 1$, the failure rate is constant and the Weibull distribution reduces to the exponential distribution. If $\gamma_1 > 1$, the the failure rate will increase with respect to time. This flexibility in failure rates make the Weibull distribution well suited to study lifetime data. For our proposed model, we will assume the pathogen promotion time to follow a Weibull distribution with density function

$$f(y) = \frac{\gamma_1}{\gamma_2}\left(\frac{y}{\gamma_2}\right)^{\gamma_1 - 1} e^{-\left(\frac{y}{\gamma_2}\right)^{\gamma_1}}, \quad y \geq 0, \gamma_1 > 0, \gamma_2 > 0. \tag{2.9}$$

Thus, we now have $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)'$. Note that one can also assume other parametric lifetime distributions here; see Balakrishnan and Pal [4], Balakrishnan and Pal [6, 7]. One can also model these promotion times assuming a semi-parametric framework (Balakrishnan et al., 2017) or assuming a completely non-parametric framework

(Balakrishnan et al., 2016). We use a completely parametric framework with respect to the pathogen promotion times. For this purpose, we define the following expressions based on our parametric Weibull assumption.

$$F_{t_k}(y|\gamma_1,\gamma_2) = \begin{cases} 1-e^{-((y-s_k)/\gamma_2)^{\gamma_1}} & ,y-s_k>0 \\ 0 & ,y-s_k\leq 0. \end{cases}$$

$$f_{t_k}(y|\gamma_1,\gamma_2) = \begin{cases} \frac{\gamma_1}{\gamma_2}\left(\frac{y-s_k}{\gamma_2}\right)^{\gamma_1-1}e^{-((y-s_k)/\gamma_2)^{\gamma_1}} & ,y-s_k>0 \\ 0 & ,y-s_k\leq 0. \end{cases}$$

$$S_{t_k}(y|\gamma_1,\gamma_2) = \begin{cases} \exp\{-((y-s_k)/\gamma_2)^{\gamma_1}\} & ,y-s_k>0 \\ 1 & ,y-s_k\leq 0. \end{cases}$$

Now, we present the explicit expressions of the Q-functions for the COM-Poisson model with multiple exposures and its special cases. We should again note that the Q-function for the COM-Poisson model and the EM algorithm using it will find the MLE's for the COM-Poisson model and all three of its special cases. However, due to the complexity of the COM-Poisson model, if we are just interested in the special cases, it is advantageous to express and use the corresponding simplified expressions for the Q-function.

2.4.1 COM-Poisson case

The Q-function, $Q(\boldsymbol{\theta},\boldsymbol{\pi}^{(r)})$, for the COM-Poisson model with a fixed dispersion parameter ν can be expressed as

$$Q(\boldsymbol{\theta},\boldsymbol{\pi}^{(r)}) = \sum_{i\in I_1}\log\left[\sum_{k=0}^{T}\left\{\frac{1}{Z(\theta_{it_k},\nu)}\frac{f_{t_k}(y_i)}{S_{t_k}(y_i)}\sum_{j=1}^{\infty}\frac{j\{\theta_{it_k}S_{t_k}(y_i)\}^j}{(j!)^v}\prod_{\substack{l\neq k\\ l=0}}^{T}\frac{Z(\theta_{it_l}S_{t_l}(y_i),\nu)}{Z(\theta_{it_l},\nu)}\right\}\right]$$
$$-\sum_{i\in I_0}(1-\pi_i^{(r)})\log\left\{\prod_{k=0}^{T}Z(\theta_{it_k},\nu)\right\}$$

$$+\sum_{i\in I_0}\pi_i^{(r)}\log\Big\{\prod_{k=0}^{T}\frac{Z(\theta_{it_k}S_{t_k}(y_i),\nu)}{Z(\theta_{it_k},\nu)}-\prod_{k=0}^{T}\frac{1}{Z(\theta_{it_k},\nu)}\Big\},$$

where

$$\pi_i^{(r)}=1-\prod_{k=0}^{T}\frac{1}{Z(\theta_{it_k}S_{t_k}(y_i),\nu)}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}}.$$

2.4.2 Poisson case

The Q-function, $Q(\boldsymbol{\theta},\boldsymbol{\pi}^{(r)})$, for the Poisson model with mean θ_{t_k} can be expressed as

$$Q(\boldsymbol{\theta},\boldsymbol{\pi}^{(r)})=\sum_{i\in I_1}\Big[\log\Big\{\sum_{k=0}^{T}\theta_{it_k}f_{t_k}(y_i)\Big\}-\sum_{k=0}^{T}\theta_{it_k}F_{t_k}(y_i)\Big]$$
$$-\sum_{i\in I_0}(1-\pi_i^{(r)})\Big\{\sum_{k=0}^{T}\theta_{it_k}\Big\}+\sum_{i\in I_0}\pi_i^{(r)}\log\Big\{e^{-\sum_{k=0}^{T}\theta_{it_k}F_{t_k}(y_i)}-e^{-\sum_{k=0}^{T}\theta_{it_k}}\Big\},$$

where

$$\pi_i^{(r)}=1-\frac{e^{-\sum_{k=0}^{T}\theta_{it_k}}}{e^{-\sum_{k=0}^{T}\theta_{it_k}F_{t_k}(y_i)}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}}.$$

2.4.3 Bernoulli case

The Q-function, $Q(\boldsymbol{\theta},\boldsymbol{\pi}^{(r)})$, for the Bernoulli model with probability of success equal to $\frac{\theta_{t_k}}{1+\theta_{t_k}}$ can be expressed as

$$Q(\boldsymbol{\theta},\boldsymbol{\pi}^{(r)})=\sum_{i\in I_1}\log\Big[\sum_{k=0}^{T}\Big[\Big\{\frac{\theta_{it_k}}{1+\theta_{it_k}}f_{t_k}(y_i)\Big\}\prod_{\substack{l\neq k\\ l=0}}^{T}\Big\{\frac{1+\theta_{it_l}S_{t_l}(y_i)}{1+\theta_{it_l}}\Big\}\Big]\Big]$$
$$-\sum_{i\in I_0}(1-\pi_i^{(r)})\Big\{\sum_{k=0}^{T}\log(1+\theta_{it_k})\Big\}$$
$$+\sum_{i\in I_0}\pi_i^{(r)}\log\Big[\prod_{k=0}^{T}\Big\{\frac{1+\theta_{it_k}S_{t_k}(y_i)}{1+\theta_{it_k}}\Big\}-\prod_{k=0}^{T}\Big\{\frac{1}{1+\theta_{it_k}}\Big\}\Big],$$

where

$$\pi_i^{(r)}=1-\prod_{k=0}^{T}\Big\{\frac{1}{1+\theta_{it_k}S_{t_k}(y_i)}\Big\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}}.$$

2.4.4 Geometric case

The Q-function $Q(\boldsymbol{\theta}, \boldsymbol{\pi}^{(r)})$, for the geometric model with parameter $1 - \theta_{t_k}$ can be expressed as

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\pi}^{(r)}) = &\sum_{i \in I_1} \log \Big[\sum_{k=0}^{T} \Big[\frac{(1-\theta_{it_k})\theta_{it_k} f_{t_k}(y_i)}{(1-\theta_{it_k} S_{t_k}(y_i))^2} \prod_{\substack{l \neq k \\ l=0}}^{T} \frac{1-\theta_{it_l}}{1-\theta_{it_l} S_{t_l}(y_i)} \Big] \Big] \\ &+ \sum_{i \in I_0} (1-\pi_i^{(r)}) \Big\{ \sum_{k=0}^{T} \log(1-\theta_{it_k}) \Big\} \\ &+ \sum_{i \in I_0} \pi_i^{(r)} \log \Big[\prod_{k=0}^{T} \frac{1-\theta_{it_k}}{1-\theta_{it_k} S_{t_k}(y_i)} - \prod_{k=0}^{T} (1-\theta_{it_k}) \Big], \end{aligned}$$

where

$$\pi_i^{(r)} = 1 - \prod_{k=0}^{T} (1-\theta_{it_k} S_{t_k}(y_i)) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}}.$$

By using these expressions of the Q-function, we can begin to study our model under different conditions using a simulation study and eventually use this model to study real data.

CHAPTER 3

DATA GENERATION

3.1 Introduction

Before a proposed model can be employed to study real data, the ability of an estimation algorithm to accurately find the maximum likelihood estimates (MLEs) of the model parameters must be tested. To do this, researchers employ a Monte Carlo simulation study. Monte Carlo simulations were conceptualized by physicist Enrico Fermi to study neutron diffusion. However, the modern version of Monte Carlo methods were developed by Stanislaw Ulam while working on the Manhattan Project, which developed the world's first atomic bomb, in Los Alamos National Laboratory. The name Monte Carlo was taken from a casino in Monaco were Ulam's uncle visited often. Monte Carlo methods work by replacing a known parameter with a probability distribution so researchers can introduce uncertainty into a model. This allows a researcher to study the effects of different yet similar conditions in their model. Returning to its original use, the scientists at Los Alamos needed a way to study the dispersion of neutrons released in the explosion of an atomic bomb. Knowing how these neutrons would propagate was impossible. So the researchers instead ran experiments where the initial parameters were randomly chosen from a specified range of possibilities, and the outcome was measured. This eventually lead to the development of the atomic bomb and the end of World War II. This new method of studying inherently random process quickly took off and became widely popular in the fields of Physics, Chemistry, Operations Research, and Economics. Most of these fields rely on Statistics, so naturally the practice became a standard in statistical research. As technology advances, Monte Carlo simulations become more accurate and intensive. In the modern era, Monte Carlo methods are used to develop new and more powerful computers utilizing machine learning and artificial intelligence.

In biostatistical research, such as this research, Monte Carlo simulations are crucial to study the effectiveness of a proposed model. To study the new model and the ability of the estimation algorithm to find the MLEs of the model parameters, we generate hundreds of data sets with known parameters and see how the new model handles the data. Is it important to see over different scenarios if the true parameter values can be accurately retrieved. This section will provide a guide to generate data for a Monte Carlo simulation study. While any programming language can be used, we used R version 4.0.0 for this simulation study. Throughout the description of the data generation, we will use the specific conditions used for our simulation study. The same process will work if the reader decides they wish to use different conditions such as a different lifetime distribution, number of patients, regression coefficients, etc.

3.2 Description of conditions

Before we describe the process of data generation, we will first describe the exact conditions used in our simulation study. To begin, this simulation study considers the pathogen promotion time of a nosocomial pulmonary infection through the use of a ventilator. Ventilators are used to help the patient breath and is employed when there is severe respiratory distress. Many of those infected with COVID-19 are forced to be placed on a ventilator due to the virus's effect on the respiratory system. This will allow us to introduce multiple exposures to a pathogen and also allow for heterogeneity in the mean number of pathogens based on exposure time. As noted in Section 2.4, we will be using the log-linear link function to model the first parameter of the COM-Poisson distribution that is related to the mean number of pathogens, i.e. θ_{t_k}. This will allow us to study heterogeneity of the infection intensity for different exposure times and for different patients. To do this, we will consider two ways a patient can be exposed to a pulmonary infection causing pathogen: through intubation and aspiration protocol. Intubation is the process of inserting a tube through the mouth of a patient and into the airway so that patient can be placed on a respirator. This is the initial moment of exposure which we will denote by t_0. To incorporate a covariate into the parameter θ_{t_0}, we will consider a binary covariate X_{imm} which represents the immunological status of the patient at the initial exposure where $X_{imm} = 0$ if the patient has a poor immunological status and =1 otherwise. Thus, θ_{t_0} at the initial exposure time will be modeled using $\theta_{t_0} = \exp(\beta_0 + X_{imm}\beta_1)$. Once the tube is inserted into the patients' airway, the doctors do not remove it until the patient has sufficiently recovered or has died. While the patient has the tube inserted in their airway, there is constant risk of aspiration, meaning the patient inhaled some foreign object, such as food, into their lungs, which can cause serious complications such as infection or tears in the airway and lungs. To prevent this, many aspiration protocols have been considered to prevent this from occurring in intubated patients. For the sake of this simulation study, we will consider two aspiration protocols: protocol A and protocol B. Since the intubation tube is not removed, let T represent the number of days the patient is intubated. Let θ_{t_k}, for $k = 1, 2, \cdots, T$, represent the infection intensity for each day of possible exposure. We can model this using $\theta_{t_k} = \exp(\beta_2 + X_{prot}\beta_3)$ where $X_{prot} = 1$ for patients undergoing protocol A and =0 otherwise. This allows us to create four groups of patients to study. Those four groups are

1. Patients with poor immunological status undergoing aspiration protocol A.

2. Patients with poor immunological status undergoing aspiration protocol B.
3. Patients with good immunological status undergoing aspiration protocol A.
4. Patients with good immunological status undergoing aspiration protocol B.

While the parameter θ_{t_k} of the COM-Poisson distribution, which models infection intensity, is heterogeneous across patients and exposure times, the dispersion parameter of the COM-Poisson distribution ν will remain the same across all patients and exposure times. Finally, we will assume the pathogen promotion time, the time from exposure to first biological sign of infection, to follow a Weibull distribution with parameters (γ_1, γ_2) as defined in (2.9). Thus, our simulation study aims to find the maximum likelihood estimate of the parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \nu)$. With this understanding, let us now discuss the method to generate artificial data for the Monte Carlo simulation.

3.3 Data generation

To begin, we must first decide the true values of the parameters used in the model. First, we need to decide the values for $(\beta_0, \beta_1, \beta_2, \beta_3)$. In this simulation study, we will assign these values as $(\beta_0, \beta_1, \beta_2, \beta_3) = (.5, -1, -3, 2)$. Next, we must decide the parameters of the Weibull distribution. We have chosen two different pairs of parameters for this study: $(\gamma_1, \gamma_2) = (2.5, 2.5)$ and $(\gamma_1, \gamma_2) = (1.5, 3.5)$, where γ_1 and γ_2 represent the shape and scale parameters, respectively, for a Weibull distribution with probability density function

$$f(y) = \frac{\gamma_1}{\gamma_2}\left(\frac{y}{\gamma_2}\right)^{\gamma_1 - 1} e^{-\left(\frac{y}{\gamma_2}\right)^{\gamma_1}}, \quad y \geq 0, \gamma_1 > 0, \gamma_2 > 0. \tag{3.1}$$

Next, we must decide on the number of patients, denoted by n, we will study and how to divide them into the four groups. We have decided to study two different scenarios: $n = 400$ divided into four equal groups of 100 and $n = 200$ divided into 4 groups of 50. These choices for sample size allows us to analyze if our model satisfies the large sample properties. Finally, we must decide the dispersion parameter of the COM-Poisson distribution. We have chosen to study the Poisson case ($\nu = 1$), the Bernoulli case ($\nu \to \infty$), and the COM-Poisson case ($\nu = 2$). As mentioned previously, due to our choice of link function between our covariates and θ_{t_k}, for $k = 0, 1, \cdots, T$, we are not guaranteed the necessary condition $0 < \theta_{t_k} < 1$ for the geometric distribution. Therefore, we cannot consider the geometric case ($\nu = 0$). Once these decisions have been made, we may proceed to the data generation. We will assume the reader is using some programming language such as R to generate these data sets.

First, we must assign the values of the parameters we have chosen for $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \nu, n)$. Once these constants have been assigned, we can generate θ_{t_k} for $k = 0, 1, \cdots, T$ using the link function we have chosen. We have saved these as a matrix for easier use later in the data generation, but the reader may wish to save them differently based on the programming language. Next we must generate random exposure times for each patient. Obviously, some patients will be intubated longer than others, so we must take this variability into account. For each patient, we generate a random number of exposure times T, in our example representing days of intubation, ranging from 2 to 30 using a discrete Uniform distribution. The reader may decide to generate the number of exposures based on another discrete distribution. Next, we must take into account the time passing between each moment of exposure. Based on the situation, this amount of time between moments of exposure may be a constant that is the same between exposures or it may be based on a continuous random variable. Depending on the data, this may represent the number of hours, days, or months between each moment of exposure. We must ensure that the amount of time added represents the amount of time from the initial exposure to each subsequent exposure time. For our purposes, we will assign the time between t_0 and t_k to be s_k, representing the number of days that have passed since the initial intubation. Now that we have the values for θ_{t_k}, number of exposures, and time between exposures, we can generate the data we will analyze. Next, we must generate censoring times for each patient. In this simulation study, random censoring times were generated from an exponential distribution with censoring rate $\alpha = .10$. Previous work such as Pal and Balakrishnan [5] have calculated exact values for the rate parameter of the exponential censoring distribution to control the overall censoring rate for their simulation study. However, due to the complexity of this model, we did not do this and instead relied solely on the random censoring rate. With these values generated, we will now describe how to generate data for each patient group.

For each group in the simulation study, the process for data generation is the same. Therefore, we will describe the data generation for a single group, which can be used for every other group. For the generation of data for each patient in a group, we need to follow these steps:

1. For each exposure time t_k, $k = 0, 1, \cdots, T$, we generate M_{t_k} pathogens from the COM-Poisson distribution with parameters θ_{t_k} and ν for a fixed value of the dispersion parameter ν. Recall θ_{t_k} may be different at the same time t_k depending on the group the patient is in.
2. We now generate random variables from the Weibull distribution to represent the pathogen promotion times.
    (a) If $M_{t_k} > 0$, we generate $\{R_{1,t_k}, R_{2,t_k}, \cdots, R_{M_{t_k},t_k}\}$ pathogen promotion times from the Weibull distribution using equation (3.1) with our chosen shape and scale parameters.

   (b) If $M_{t_k} = 0$, then $R_{0,t_k} = \infty$.

3. We must account for the time between each moment of exposure.
   (a) Let s_k denote the time between the initial exposure time, t_0, and the k^{th} exposure time t_k.
   (b) Let $Z_{i,t_k} = R_{i,t_k} + s_k$ for $k = 1, 2, \cdots, T$ and $i = 1, 2, \cdots, M_{t_k}$, which represents the re-parameterized time-scale of infection as covered in Section 2.3.
   (c) For $k = 0$, $Z_{i,t_0} = R_{i,t_0}$.
4. Let $W_{t_k} = \min\{Z_{1,t_k}, Z_{2,t_k}, \cdots, Z_{M_{t_k},t_k}\}$, which represents the time-to-event at exposure time t_k. If $M_{t_k} = 0$, we let $W_{t_k} = \infty$.
5. Since we have competing pathogens at multiple exposure times, we define $W = \min\{W_{t_0}, W_{t_1}, \cdots, W_{t_k}, \cdots, W_{t_T}\}$ as the true time-to-event. Furthermore, let Y denote the observed time-to-event.
6. Let C denote the random censoring time generated form an exponential distribution with a suitable rate α to meet a desired censoring proportion.
   (a) If $W = \infty$, that is, the patient is exposed to no pathogens across all exposure times, meaning the patient is immune to the event of interest, we set $Y = C$.
   (b) If $W < \infty$, we set $Y = \min\{W, C\}$.
7. Let δ denote the binary right censoring indicator.
   (a) If $Y = W$, we set $\delta = 1$, meaning the true lifetime is observed.
   (b) If $Y = C$, we set $\delta = 0$, meaning the lifetime is right censored.

We now have the desired ordered pair of values (Y, δ) representing the lifetime and censoring indicator we need to find the maximum likelihood estimates of the model parameters. Next we must create two functions for the simulation study. Recall the long-term density function for the COM-Poisson distribution is

$$\sum_{k=0}^{T}\left[\frac{1}{Z(\theta_{t_k},\nu)}\frac{f_{t_k}(y)}{S_{t_k}(y)}\sum_{j=1}^{\infty}\frac{j\{\theta_{t_k}S_{t_k}(y)\}^j}{(j!)^v}\prod_{\substack{l\neq k\\ l=0}}^{T}\frac{Z(\theta_{t_l}S_{t_l}(y),\nu)}{Z(\theta_{t_l},\nu)}\right].$$

Note that $Z(\cdot,\cdot)$ is the normalizing constant in the COM-Poisson distribution calculated by

$$Z(\theta,\nu)=\sum_{j=0}^{\infty}\frac{\theta^j}{(j!)^{\nu}}.$$

We need to create functions to handle the infinite series $Z(\cdot,\cdot)$ and $\sum_{j=1}^{\infty}\frac{j\{\theta_{t_k}S_{t_k}(y)\}^j}{(j!)^v}$. For both these infinite series, we create functions that calculate these infinite series up to 50 terms. While this does not give us the exact value of these functions, it does give us a good approximation of the true value. After 50 terms, the new terms are so small, R and many other programming languages cannot calculate them. Therefore, we truncate the infinite series after 50 terms. Finally, we must re-parameterize the lifetimes, censored and observed, with the time-scale of infection on the original scale to use for the re-parameterized survival function $S_{t_k}(y|\boldsymbol{\gamma})$ and re-parameterized cumulative distribution function $F_{t_k}(y|\boldsymbol{\gamma})$. Let (Y_i,δ_i) denote the lifetime and censoring indicator as defined previously for the i^{th} patient for $i=1,\cdots,n$. Let the i^{th} patient be exposed to T_i many moments of infection. Then, let $y_{i,k}$ be defined as

$$y_{i,k}=\begin{cases}Y_i-s_k, & y-s_k>0\\ 0, & y-s_k\leq 0.\end{cases}\tag{3.2}$$

for $k=1,\cdots,T_i$. These new re-parameterized lifetimes will be used for as the inputs for F_{t_k} and S_{t_k}. In other words, $y_{i,k}$ will be used as the input for F_{t_k} and S_{t_k} for $k=1,\cdots,T_i$.

CHAPTER 4

SIMULATION STUDY

We will now present the results of the simulation study conducted on our proposed model using the framework for data generation described in Chapter 3.

4.1 Model fitting

4.1.1 Special cases

As described in Chapter 3, we will consider four different combinations of sample size and model parameters to study during this simulation study

1. $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, n) = (.5, -1, -3, 2, 2.5, 2.5, 400)$.
2. $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, n) = (.5, -1, -3, 2, 2.5, 2.5, 200)$.
3. $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, n) = (.5, -1, -3, 2, 1.5, 3.5, 400)$.
4. $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, n) = (.5, -1, -3, 2, 1.5, 3.5, 200)$.

Recall we will not be analyzing the geometric special case of the COM-Poisson distribution, when $\nu = 0$, since our chosen link function does not guarantee $0 < \theta_{t_k} < 1$ for all $k = 0, 1, \cdots, T$. To evaluate the performance of the estimates of the parameters, we calculated the bias and root mean square errors (RMSE) of all the estimates. Additionally, we calculated the 95% coverage probabilities (CP) of the confidence intervals based on the asymptotic normality of the MLEs. All of these simulations were performed using R software version 4.0.0 and the results were based on 200 Monte Carlo runs. While a typical Monte Carlo simulation study uses more simulations, the we decided 200 Monte Carlo simulations were sufficient since the complexity of the model requires a lot of computational power which in turn takes a very long time to run.

The EM algorithm requires an initial guess of the parameters to begin the iterative process. To find our initial guess for each parameter, we used a random number from a uniform distribution whose minimum and maximum values were a 15% deviation from the true value of the parameter on either side. This ensured different initial guesses for each Monte Carlo simulation. This randomness allowed us to explore the efficiency of our model with different generated data and different initial guesses.

To calculate the asymptotic standard error of each parameter, we need to find the inverse of the Hessian matrix. The asymptotic standard errors can be found by calculating the inverse of the hessian matrix at the MLE values and taking the square root of the diagonal values. While it is possible to derive the Hessian matrix by hand, this is not ideal due to the complexity of the model. Therefore, we have used the R package "numDeriv" to calculate the standard errors. Within this package, there is a function "hessmat" which provides an accurate numerical approximation for the Hessian matrix at the given values. We have taken advantage of this function for efficient calculation of the asymptotic standard errors. Once these standard errors have been calculated, we can easily find the coverage probability of each parameter.

Tables 4.1 and 4.2 show the results of our Monte Carlo simulation study for the Bernoulli and Poisson models, respectively. It is clear to see that the EM algorithm can accurately retrieve the true parameters of the underlying model. The bias, standard error, and RMSE are all relatively small which is desired in model fitting. Furthermore, the 95% coverage probabilities are all close to nominal level. The tables also show our model obeys the large sample theory since the standard error, bias, and RMSE all decrease as the sample size increases. While these results show a great ability of our proposed method to find the true parameters of the model, it does so with greater accuracy for the models in which $(\gamma_1, \gamma_2) = (2.5, 2.5)$. While the estimates for the cases when $(\gamma_1, \gamma_2) = (1.5, 3.5)$ are still good, the estimates for the other case are certainly better.

4.1.2 General case of COM-Poisson model

Let us now discuss the parameter estimation of the general COM-Poisson model. While the EM algorithm does a very good job predicting the true parameters for the covariates and the promotion time distribution parameters, it is not as well equipped to estimate the dispersion parameter ν of the COM-Poisson distribution. As such, a separate technique will be needed in conjunction with the EM algorithm. To estimate ν, we will employ a profile likelihood technique within the EM algorithm, which is done along the lines of Balakrishnan and Pal [8]. For this purpose, we first select a set of admissible values for the dispersion parameter ν. Then, for each of these chosen values, we run the EM algorithm using the chosen value of ν as a constant in the model to estimate the other parameters in the model. We next calculate the log-likelihood value at the MLEs. Finally, we select the MLE of ν as the value of ν that results in the highest log-likelihood value. We use that value of ν and the MLEs of the other parameters to evaluate the general COM-Poison model. In this simulation study for the general COM-Poisson model, we will use the same four settings that were used in the special cases models. We have chosen the true parameter of ν as 2, which simulates data that is under-dispersed relative to the Poisson distribution. For the admissible values of ν, we have chosen $\{1.6, 1.7, \cdots, 2.3, 2.4\}$ which gives us eleven possible choices of ν. The results of this simulation study can be found in Table 4.3. Similar to the results for the Poisson and Bernoulli models, the general EM algorithm estimates the COM-Poisson model parameters quite accurately. The bias, standard error, and RMSE are all relatively small. The 95% coverage probabilities are all close to the nominal level, and all the large sample properties are satisfied as well.

Table 4.1. Model fitting results for the Bernoulli cure rate model with multiple exposures.

n	True Value	Estimate	S.E.	Bias	RMSE	CP
400 $(100, 100, 100, 100)$	$\beta_0 = 0.5$	0.4770	0.3005	-0.0230	0.2898	0.950
	$\beta_1 = -1$	-1.0438	0.3191	-0.0438	0.3225	0.940
	$\beta_2 = -3$	-2.9912	0.2890	0.0088	0.2619	0.940
	$\beta_3 = 2$	2.0321	0.3336	0.0321	0.2695	0.970
	$\gamma_1 = 2.5$	2.5554	0.2095	0.0554	0.2213	0.935
	$\gamma_2 = 2.5$	2.4398	0.1608	0.0602	0.1457	0.965
200 $(50, 50, 50, 50)$	$\beta_0 = 0.5$	0.5270	0.4168	0.0270	0.3982	0.940
	$\beta_1 = -1$	-1.0379	0.4649	-0.0379	0.4955	0.935
	$\beta_2 = -3$	-3.1321	0.4935	-0.1321	0.7061	0.960
	$\beta_3 = 2$	2.2151	0.5552	0.2151	0.7774	0.930
	$\gamma_1 = 2.5$	2.5889	0.2852	0.0889	0.3335	0.925
	$\gamma_2 = 2.5$	2.4983	0.2096	0.0017	0.1948	0.940
400 $(100, 100, 100, 100)$	$\beta_0 = 0.5$	0.4155	0.4383	-0.0845	0.4395	0.950
	$\beta_1 = -1$	-1.0541	0.3935	-0.0541	0.3331	0.975
	$\beta_2 = -3$	-2.9738	0.3831	0.0262	0.3973	0.930
	$\beta_3 = 2$	2.9755	0.4174	-0.0245	0.4588	0.915
	$\gamma_1 = 1.5$	1.5570	0.1274	0.0570	0.1359	0.965
	$\gamma_2 = 3.5$	3.3984	0.3799	-0.1016	0.4136	0.905
200 $(50, 50, 50, 50)$	$\beta_0 = 0.5$	0.4308	0.6067	-0.0692	0.5680	0.935
	$\beta_1 = -1$	-0.9745	0.5598	0.0255	0.5332	0.950
	$\beta_2 = -3$	-2.9882	0.5582	0.0118	0.4703	0.920
	$\beta_3 = 2$	1.9906	0.6054	-0.0094	0.5151	0.910
	$\gamma_1 = 1.5$	1.5788	0.1911	0.0788	0.2375	0.935
	$\gamma_2 = 3.5$	3.3074	0.5371	-0.1916	0.5426	0.895

Table 4.2. Model fitting results for the Poisson cure rate model with multiple exposures.

n	True Value	Estimate	S.E.	Bias	RMSE	CP
400 $(100, 100, 100, 100)$	$\beta_0 = 0.5$	0.5060	0.1262	0.0060	0.1239	0.955
	$\beta_1 = -1$	-1.0137	0.1605	-0.0137	0.1757	0.935
	$\beta_2 = -3$	-3.0621	0.4553	-0.0621	0.4644	0.960
	$\beta_3 = 2$	2.0614	0.4936	0.0614	0.5037	0.965
	$\gamma_1 = 2.5$	2.5282	0.1654	0.0282	0.1687	0.960
	$\gamma_2 = 2.5$	2.4936	0.1604	-0.0064	0.1577	0.965
200 $(50, 50, 50, 50)$	$\beta_0 = 0.5$	0.5110	0.1785	0.0110	0.1713	0.970
	$\beta_1 = -1$	-1.0370	0.2303	-0.0370	0.2326	0.965
	$\beta_2 = -3$	-3.0768	0.4916	-0.0768	0.4923	0.965
	$\beta_3 = 2$	2.0866	0.5459	0.0866	0.5317	0.955
	$\gamma_1 = 2.5$	2.5326	0.2329	0.0326	0.3536	0.945
	$\gamma_2 = 2.5$	2.5180	0.2264	0.0180	0.2318	0.930
400 $(100, 100, 100, 100)$	$\beta_0 = 0.5$	0.5069	0.1867	0.0069	0.1818	0.960
	$\beta_1 = -1$	-1.1014	0.1817	-0.1014	0.1823	0.945
	$\beta_2 = -3$	-3.1269	0.6654	-0.1269	0.6932	0.900
	$\beta_3 = 2$	2.1127	0.6983	0.1127	0.7281	0.910
	$\gamma_1 = 1.5$	1.5155	0.1062	0.0155	0.1033	0.965
	$\gamma_2 = 3.5$	3.6233	0.4127	0.1233	0.4266	0.905
200 $(50, 50, 50, 50)$	$\beta_0 = 0.5$	0.4729	0.2687	-0.0281	0.2744	0.945
	$\beta_1 = -1$	-1.0477	0.2688	0.0477	0.2653	0.965
	$\beta_2 = -3$	-3.0373	0.7598	-0.0373	0.5923	0.920
	$\beta_3 = 2$	2.0541	0.8183	0.0541	0.6590	0.915
	$\gamma_1 = 1.5$	1.5466	0.1579	0.0466	0.1769	0.930
	$\gamma_2 = 3.5$	3.4580	0.8309	0.0420	0.8945	0.905

Table 4.3. Model fitting results for the COM-Poisson cure rate model with multiple exposures (to apply the profile likelihood, the set of values of ν is chosen as $\{1.6, 1.7, \cdots, 2.4\}$).

n	True Value	Estimate	S.E.	Bias	RMSE	CP
400 $(100, 100, 100, 100)$	$\beta_0 = 0.5$	0.4644	0.2259	-0.0356	0.2282	0.945
	$\beta_1 = -1$	-1.0151	0.2345	-0.0151	0.2344	0.940
	$\beta_2 = -3$	-3.0504	0.2121	-0.0504	0.2175	0.960
	$\beta_3 = 2$	2.0522	0.2177	0.0522	0.2233	0.960
	$\gamma_1 = 2.5$	2.5842	0.2575	0.0842	0.2703	0.965
	$\gamma_2 = 2.5$	2.5376	0.2569	0.0376	0.2590	0.955
	$\nu = 2$	1.8320	–	-0.1680	0.8133	–
200 $(50, 50, 50, 50)$	$\beta_0 = 0.5$	0.4554	0.5288	-0.0446	0.5280	0.940
	$\beta_1 = -1$	-1.1001	0.4805	-0.1001	0.5149	0.950
	$\beta_2 = -3$	-3.0033	0.3293	-0.0033	0.3277	0.960
	$\beta_3 = 2$	1.9846	0.3507	-0.0126	0.3492	0.960
	$\gamma_1 = 2.5$	2.6991	0.4356	0.1991	0.4770	0.925
	$\gamma_2 = 2.5$	2.5084	0.3260	0.0084	0.3345	0.920
	$\nu = 2$	1.7975	–	-0.2025	0.8956	–
400 $(100, 100, 100, 100)$	$\beta_0 = 0.5$	0.4470	0.3462	-0.0630	0.3485	0.950
	$\beta_1 = -1$	-1.0812	0.4155	-0.0812	0.4213	0.945
	$\beta_2 = -3$	-3.0359	0.2481	-0.0359	0.2494	0.910
	$\beta_3 = 2$	2.0275	0.2792	0.0275	0.2791	0.920
	$\gamma_1 = 1.5$	1.5556	0.1629	0.0556	0.1714	0.945
	$\gamma_2 = 3.5$	3.6636	0.5117	0.1636	0.5348	0.950
	$\nu = 2$	2.1878	–	0.1878	0.8401	–
200 $(50, 50, 50, 50)$	$\beta_0 = 0.5$	0.5433	0.5760	0.0433	0.5746	0.925
	$\beta_1 = -1$	-1.1837	0.6957	-0.1837	0.7579	0.945
	$\beta_2 = -3$	-2.9712	0.3775	0.0288	0.3766	0.900
	$\beta_3 = 2$	1.9377	0.3555	-0.0623	0.3591	0.915
	$\gamma_1 = 1.5$	1.5665	0.2647	0.0665	0.2716	0.935
	$\gamma_2 = 3.5$	3.5186	0.7116	0.0186	0.7082	0.940
	$\nu = 2$	1.7864	–	-0.2136	0.9376	–

4.2 Model discrimination

Due to the flexibility of the multiple exposure COM-Poisson model and its inclusion of several other multiple exposure models as special cases, we are in a position to select a simple multiple exposure model within the bigger family of COM-Poisson multiple exposure model that provides an adequate fit as the COM-Poisson model itself in many cases. This motivates us to explore the flexibility of the multiple exposure COM-Poisson cure rate model to select a parsimonious cure rate model that provides an adequate fit to the given data. To this end, we conduct two different model discrimination studies, one using the likelihood ratio test and the other using the information-based criteria.

4.2.1 Likelihood ratio test

In this model discrimination study, we investigate the performance of the likelihood ratio test to test the null hypothesis that the distribution of the number of pathogens at each exposure time can be described by one of the Bernoulli ($\nu \to \infty$), Poisson ($H_0 : \nu = 1$), COM-Poisson ($\nu = 0.5$), and COM-Poisson ($\nu = 2$) distributions versus the alternative hypothesis that the number of pathogens can be described by any other member of the COM-Poisson family besides the one already specified in the null hypothesis. The likelihood test statistic is defined as $\Lambda = -2(\hat{l}_0 - \hat{l})$, where $\hat{l}_0$ denotes the maximized log-likelihood function value under the null hypothesis and $\hat{l}$ denotes the unrestricted maximized log-likelihood function value. Note that to calculate $\hat{l}$, we fit the multiple exposure COM-Poisson model for which the profile likelihood technique needs to be employed. For this simulation study, we consider the following three parameter settings: (i) Setting 1 considers 400 patients with parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2) = (0.5, -1, -3, 2, 2.5, 2.5)$; (ii) Setting 2 considers 400 patients with parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2) = (0.5, -1, -3, 2, 1.5, 3.5)$; and (iii) Setting 3 considers 200 patients with parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2) = (0.5, -1, -3, 2, 2.5, 2.5)$. For each simulated data from a true model, we calculate the likelihood ratio test statistic of the fitted Bernoulli, Poisson, COM-Poisson ($\nu = 0.5$), and COM-Poisson ($\nu = 2$) models versus the fitted COM-Poisson model. Based on 200 data sets for each true model and for each parameter setting, and using 10% level of significance, we report the observed significance levels (in bold) and observed power values of the likelihood ratio test in Table 4.4. These values are obtained by the rejection rates of the null hypothesis. From Table 4.4, it is clear that the asymptotic null distribution of the likelihood ratio test statistic is reasonably approximated with all observed levels being above the true nominal level of 10%. The approximation only turns out to be better for the Bernoulli cure rate model, in which case the observed levels are close to the true level. When the true model is Bernoulli (or COM-Poisson ($\nu = 0.5$)), the power to reject the COM-Poisson ($\nu = 0.5$) (or Bernoulli) is high. Thus, the likelihood ratio test can discriminate between the Bernoulli and COM-Poisson ($\nu = 0.5$) models. In this regard, note that the rejection rate is higher when the true model is COM-Poisson ($\nu = 0.5$) and the fitted model is Bernoulli. Now, when the true model is Poisson, the likelihood ratio test still possess adequate power to reject the Bernoulli model. However, when the true model is Bernoulli, the test has very low power to reject the Poisson model. Finally, the power of the likelihood ratio test to discriminate among COM-Poisson ($\nu = 0.5$), Poisson, and COM-Poisson ($\nu = 2$) models vary from low to medium.

4.2.2 Information-based criteria

In this model discrimination study, we investigate the performance of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) in choosing either the Bernoulli multiple exposure model, Poisson multiple exposure model, COM-Poisson ($\nu = 0.5$) multiple exposure model or COM-Poisson ($\nu = 2$) multiple exposure model, for a given true multiple exposure model. Our selected models cover both over-dispersed and under-dispersed models. We choose to look into the AIC and BIC since they are they two most widely used model selection criteria. The AIC was first introduced by Hirotugu Akaike (see Akaike [1]) and is

Table 4.4. Observed levels and observed power values of the likelihood ratio test.

	True multiple exposure model			
Fitted model	$\nu = 0.5$	$\nu = 1$	$\nu = 2$	$\nu \to \infty$
		Setting 1		
$\nu = 0.5$	**0.100**	0.315	0.465	0.765
$\nu = 1$	0.230	**0.095**	0.105	0.140
$\nu = 2$	0.425	0.390	**0.100**	0.095
$\nu \to \infty$	0.945	0.745	0.400	**0.070**
		Setting 2		
$\nu = 0.5$	**0.115**	0.250	0.400	0.695
$\nu = 1$	0.205	**0.125**	0.085	0.155
$\nu = 2$	0.435	0.350	**0.160**	0.105
$\nu \to \infty$	0.885	0.715	0.410	**0.085**
		Setting 3		
$\nu = 0.5$	**0.145**	0.225	0.380	0.680
$\nu = 1$	0.180	**0.150**	0.075	0.125
$\nu = 2$	0.375	0.355	**0.170**	0.090
$\nu \to \infty$	0.845	0.650	0.395	**0.065**

defined as $AIC = -2l + 2p$, where l is the maximized log-likelihood value of the given model and p is the number of parameters of the fitted model. The BIC, on the other hand, was developed by Gideon Schwarz (see Schwarz [50]) and is defined as $BIC = -2l + p\log(n)$. Similar to the AIC, l is the maximized log-likelihood value of the given model, p is the number of parameters of the fitted model, and n is the sample size. In both AIC and BIC, the preferred model is the fitted model with the lowest AIC or BIC value.

For each generated true model, we fit all candidate models, i.e., Bernoulli ($\nu \to \infty$), Poisson ($\nu = 1$), COM-Poisson ($\nu = 0.5$), and COM-Poisson ($\nu = 2$), and allow AIC/BIC to select the best model. To generate the true model, we consider the same parameter settings as in the case of likelihood ratio test. Based on 200 generated data sets for each true model and for each parameter setting, we calculate the observed selection rates for both AIC and BIC, and report these values in Table 4.5. Note that the selection rates for BIC turned out to be the same as that for AIC, and, as such, are not reported. From the results in Table 4.5, it is clear that the model selection criteria performs well in selecting the correct model. When the true model is Bernoulli (or COM-Poisson ($\nu = 0.5$)), the selection rate for COM-Poisson ($\nu = 0.5$) (or Bernoulli) is very low. This suggests that the AIC and BIC can distinctly discriminate between these two models. Thus, the decision reached by AIC/BIC is in agreement with that reached by likelihood ratio test earlier. A similar conclusion can also be drawn when discriminating between Bernoulli and COM-Poisson ($\nu = 2$) as well as between Bernoulli and Poisson models. However, the discrimination power of AIC and BIC among COM-Poisson ($\nu = 0.5$), Poisson and COM-Poisson ($\nu = 2$) models appear to be weak. Note, in this regard, that when the true model is COM-Poisson ($\nu = 2$), the selection rate for COM-Poisson ($\nu = 0.5$) is low. These observations suggest that it is worth exploring the flexibility of the COM-Poisson distribution to select a suitable distribution for the count on the pathogens at each exposure time. We note the advantage of using AIC/BIC in terms of speed. Unlike the method of model discrimination using likelihood ratio test, we do not need to estimate the shape parameter v since we are only attempting to fit the candidate models for which the values of ν are specified. If we were to estimate the shape parameter ν, we would need to use a profile likelihood approach, which would drastically increase the computation time.

From these results, we can see our proposed EM algorithm can accurately predict the true values of the underlying model with relatively low bias, standard error, and root mean square error. Furthermore, the likelihood ratio test and the AIC/BIC are able to distinguish between different cases of the COM-Poisson cure rate model quite well. As such, we can now apply this model and the estimation

Table 4.5. Observed selection rates based on AIC

	True multiple exposure model			
Fitted model	$\nu = 0.5$	$\nu = 1$	$\nu = 2$	$\nu \to \infty$
		Setting 1		
$\nu = 0.5$	0.405	0.275	0.150	0.035
$\nu = 1$	0.310	0.410	0.315	0.105
$\nu = 2$	0.235	0.215	0.330	0.145
$\nu \to \infty$	0.050	0.100	0.205	0.715
		Setting 2		
$\nu = 0.5$	0.415	0.290	0.165	0.060
$\nu = 1$	0.320	0.425	0.320	0.110
$\nu = 2$	0.215	0.200	0.325	0.125
$\nu \to \infty$	0.050	0.085	0.190	0.705
		Setting 3		
$\nu = 0.5$	0.420	0.285	0.140	0.025
$\nu = 1$	0.300	0.400	0.295	0.140
$\nu = 2$	0.235	0.220	0.340	0.165
$\nu \to \infty$	0.045	0.095	0.225	0.670

algorithm to a real data set to study the underlying conditions of the model. This will allow us to study infectious diseases with more accuracy.

CHAPTER 5

REAL DATA ANALYSIS

5.1 Description of data

Now that we have demonstrated our proposed algorithm's ability to accurately predict the maximum likelihood estimates and find the underlying distribution for the model, we can use our proposed model and algorithm to study real data. The data used in this study can be found via the website *Kaggle*. *Kaggle* is a free-to-access online database repository that contains thousands of data sets created by researchers worldwide. The data we will be evaluating was collected by the World Health Organization (WHO) in conjunction with John Hopkins University on December 31, 2019. This data tracks patients who have contracted the SARS-CoV-2, also known as COVID-19 or simply coronavirus, after visiting Wuhan City, Hubei Province of China, which is believed to the epicenter of the COVID-19 pandemic. In 2020, COVID-19 has been a focus of medical research due to the virus' infection rate, world-wide spread, and mortality rate. The virus is easily spread and causes respiratory distress in some patients. Some patients experience severe respiratory symptoms which can lead to death and others are asymptomatic carriers of the virus. This data set contains the patient data of people who have been in contact with people from Wuhan or someone who had recently visited Wuhan. The data set has information on variables such as country of origin, gender, age, date of symptom onset, if the patient died, and if the patient recovered. Most importantly, the data set contains the day when the patient started to be exposed to COVID-19 and when the exposure ended, which is crucial for our model. However, not all patients had this information. As such, we had to comb through the data to find patients who had the values we desired, so we could create a new data set with the information we required. We began by selecting patients who had a definitive exposure start date and exposure end date. This left us with 120 patients to study. We decided the event of interest we wanted to study would be the time to recovery. From these patients, we decided to study the effects of the covariates age and gender on the recovery time of the patients. Once we eliminated patients whose age and gender were not available, we were left with 95 patients whose starting exposure time, ending exposure time, age, and gender were recorded. Of these 95 patients, 15 of them had recovered from the disease and their date of recovery was also recorded.

5.2 Data preparation for model

Now that we have our desired data set of 95 patients, we can use our proposed model and algorithm to find the MLEs of the model. To begin, we identify our 2 covariates of interest as age and gender denoted X_{age} and X_{gen}, respectively. We will treat age as a binary covariate where $X_{gen} = 1$ if the patient is male (54.7%) and $X_{gen} = 0$ if the patient is female (45.3%). Next, we will let our covariate age be a categorical covariate. It would be impractical to divide the patients into groups based on their exact age, so instead we broke the patients into 5 groups that were representative of the dispersion of the age. $X_{age} = 1$ if the patient is between 0 and 24.5 years old (8.4%), $X_{age} = 2$ if the patient is between 24.5 and 36.5 years old (23.2%), $X_{age} = 3$ if the patient is between 36.5 and 48.5 years old (29.5%), $X_{age} = 4$ if the patient is between 48.5 and 60.5 years old (25.3%), and $X_{age} = 5$ if the patient is older than 60.5 years old (13.7%). Now that we have defined our covariates and their values, we can use the log linear link function as $\theta_{t_k} = exp(\beta_0 + \beta_1 X_{gen} + \beta_2 X_{age})$. Since we have no evidence of heterogeneity of exposure intensity with respect to moment of exposure, we will assume θ_{t_k} is the same for each moment of exposure for each patient. Since the data records the number of days each patient was exposed to COVID-19, we will let the number of exposures be the number of days the patient was exposed. Therefore, the “time jump” between each moment of exposure as explained previously in this dissertation will be 1. The average number of exposures is 6.24, with a minimum value of 1 and a maximum value of 29. Furthermore, we will let the value for the days until recovery, in other words the time to the event of interest, to be the number of days that have passed from the first day of exposure until the recovery date. The average time to recovery is 29.125 days, the minimum value is 22 days, and the maximum value is 40 days. If the recovery time of the patient is not recorded, the patient is censored and we allow the lifetime value to we will discuss to be 40 days. Next, to begin the iterative process of the EM algorithm, an initial guess of parameters is needed. To find the initial guess of the parameters in the Weibull distribution, we found the mean and variance of the recorded lifetimes. We then used the known expressions for the mean and variance of the Weibull distribution and solved for the progression time parameters. To get an initial guess for the regression coefficients, we performed a grid search using the observed likelihood function. For the purpose of the grid search, we assumed the data to follow the Poisson distribution. We then set the values for the lifetime parameters as constant. Finally, we calculated the log-likelihood values using the constant lifetime parameters, and different combinations of the regression coefficients with values ranging from $[-5, 5]$. Once we found the maximum log-likelihood value, we used those chosen parameters as the initial guess to begin the iterative process of the COM-Poisson distribution. We used these initial guesses and different values of ν to determine the correct model for this data. We selected values of ν ranging from 0 to 2, with jumps of .1, as well as the Bernoulli case ($\nu \to \infty$). Now that the data has been prepared, we can show the results of our study.

5.3 Real data results

Our model was able to converge to values for the regression coefficients and lifetime parameters very well. The algorithm was able to converge to the same values of $(\beta_0, \beta_1, \beta_2, \gamma_1, \gamma_2)$ very quickly and under different, yet relatively close, values of the starting guesses. However, the model was not able to accurately distinguish between the different values of ν. Figure 5.1 shows the value of the maximized log-likelihood function using our selected values of ν. As we can clearly see, the values of the maximized log-likelihood function are very close to each other and almost indistinguishable. Furthermore, Table 5.1 shows some selected values of ν from the real data analysis along with the maximized log-likelihood , AIC, and BIC values. From Table 5.1, we can see there is very little difference between the different values of ν. Therefore, we can say that any model within the COM-Poisson family is adequate for this data. Therefore, we can assume our model follows any of our proposed models. For the remainder of this chapter, we will assume the data follows the geometric cure rate model since it has the highest log-likelihood value and lowest AIC and BIC values. Under this assumption, we can look at the Kaplan Meier curves to study the effect the covariates have on the overall survival of the population. Figure 5.2 shows the Kaplan Meier curve that demonstrates the effect of the covariate “age” on the survival of the male population. Figure 5.3 shows the Kaplan Meier curve that demonstrates the effect of the covariate “age” on the survival of the female population.

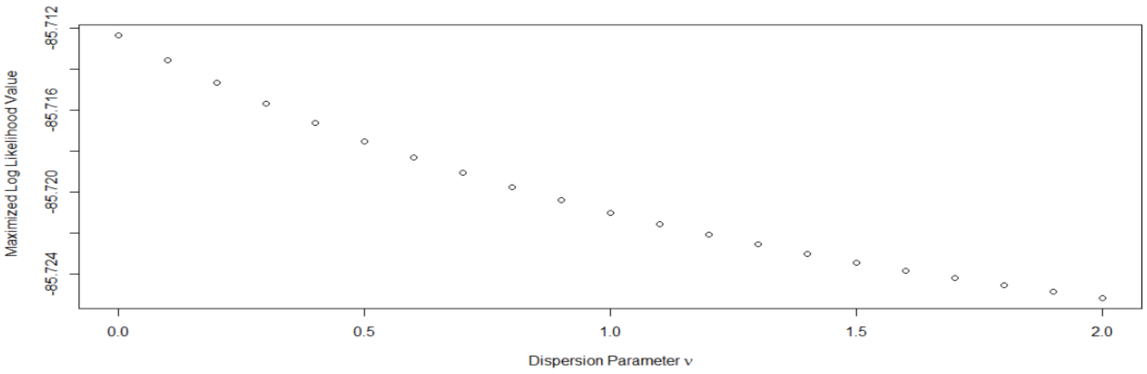

Figure 5.1. *This figure shows plots the selected values of ν against the maximized log-likelihood values. .*

Table 5.1. AIC, BIC and maximized log-likelihood function ($\hat{l}$) values for different cure rate models.

Model	$\hat{l}$	AIC	BIC
COM-Poisson (geometric)	-85.7124	183.4247	198.7480
COM-Poisson ($\nu = 0.5$)	-85.7175	183.4350	198.7583
COM-Poisson (Poisson)	-85.7210	183.4420	198.7653
COM-Poisson ($\nu = 2$)	-85.7252	183.4503	198.7736
COM-Poisson (Bernoulli)	-85.7294	183.4588	198.7820

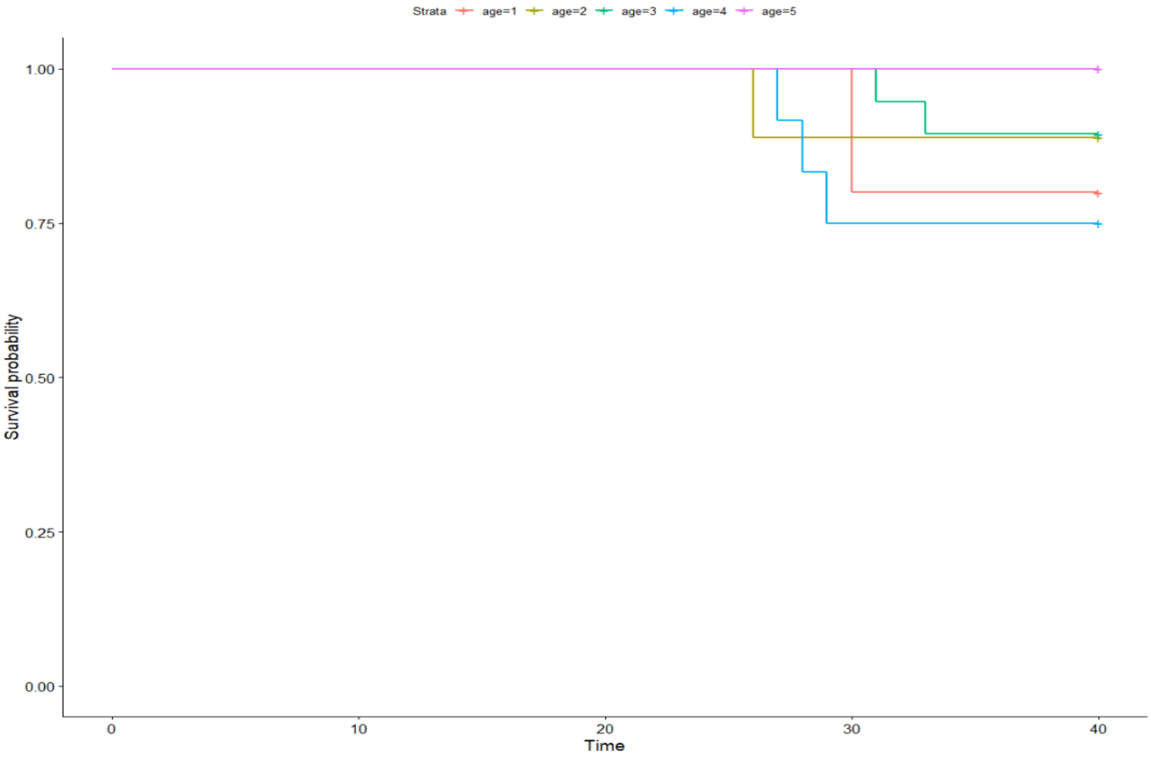

Figure 5.2. *This figure shows the Kaplan Meier curve to show the effect of age if the patient is male .*

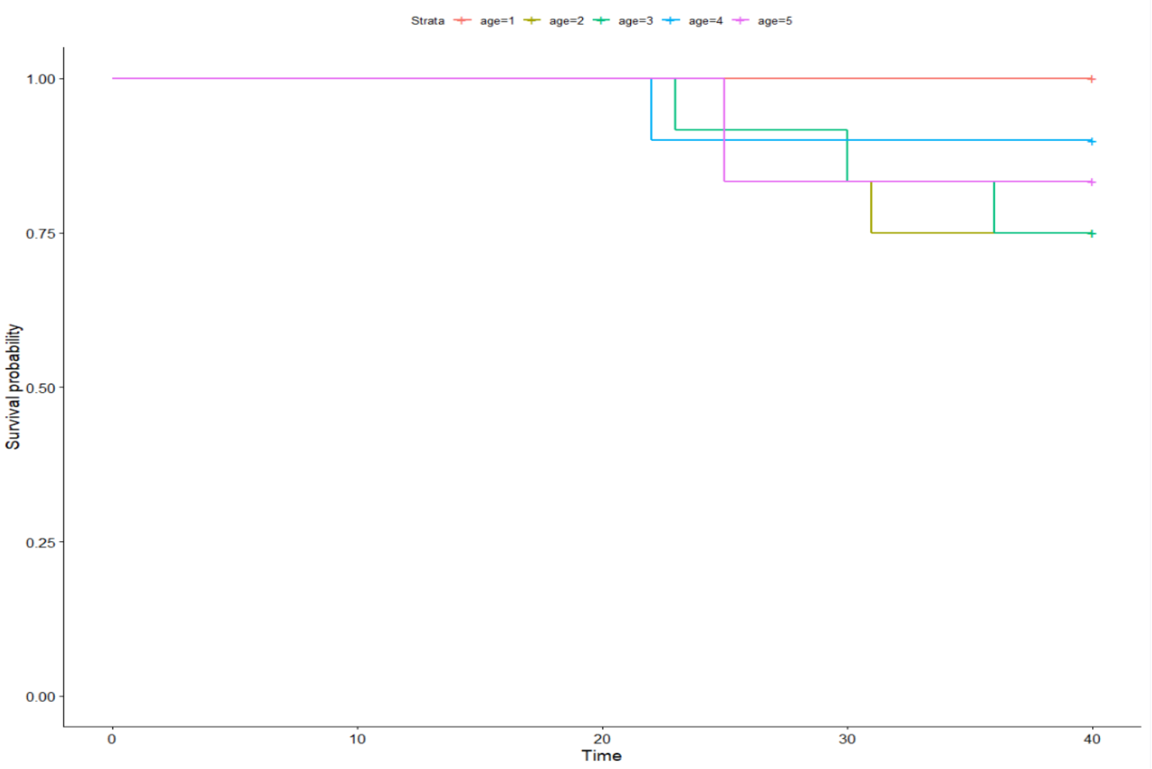

Figure 5.3. *This figure shows the Kaplan Meier curve to show the effect of age if the patient is female .*

CHAPTER 6

SUMMARY OF RESEARCH

When studying most infectious diseases, some patients exposed to the disease of interest will be asymptomatic, show no biological signs of the disease, or may be all together immune to the disease. Others infected may succumb to the disease more rapidly than others. It is of upmost importance to those studying the infectious disease to accurately estimate the proportion of the population immune or asymptomatic to the disease and to find underlying factors that can accelerate the progression of the pathogen. Studying both of these populations and understanding the factors that contribute to both groups are vital in slowing down or stopping the spread of an infectious disease. In an era where the threat of antibiotic resistant bacteria, so called "super bugs", grows, understanding new statistical methods to slow the progression will become more important.

Tournoud and Ecochard [55] developed a cure rate model for infectious diseases that allowed for multiple and discrete exposures to the disease. In their model, they assumed the number of competing causes to follow the Poisson distribution. This work was later expanded upon by Tournoud and Ecochard [56] in which they used the Poisson, Bernoulli, negative binomial, and compound Poisson distribution to model the number of competing pathogens at each moment of exposure. However, as mentioned in Section 2.3, the compound Poisson distribution proposed by Tournoud and Ecochard [56] does not have a clear biological interpretation for modeling the number of competing causes as it is a discrete distribution derived from the sum of continuous random variables. In this thesis, we have considered the cure rate model with multiple exposures developed by Tournoud and Ecochard [55]. The main contribution of this work is the introduction of the COM-Poisson distribution to address the issue with the compound Poisson distribution and developing an exact EM algorithm for estimating the parameters of the COM-Poisson cure rate model with multiple exposures and its special cases. This introduces a new model with far more flexibility and natural biological interpretation.

In this thesis, we considered the lifetime distribution modeling the pathogen promotion times to follow a Weibull distribution. We then developed the necessary steps of the EM algorithm to find the MLEs of the parameters of our proposed model. Two studies were carried out with the use of Monte Carlo simulations; one dealing with the estimation of the parameters and evaluating the performance of the proposed method of finding the MLEs, and the other demonstrating the flexibility of the COM-Poisson family to select a proper competing pathogen distribution that provides an adequate fit to the data. The results shown in Chapter 4 demonstrate our proposed methodology's ability to find estimates that converge to the true parameters of the model quite accurately. Since the likelihood surface turned out to be flat with respect to the dispersion parameter of the COM-Poisson distribution, a profile likelihood approach was used to estimate ν. For the model discrimination study, we used the AIC and BIC as well as the likelihood ratio test. When investigating the ability of the AIC and BIC to discriminate between models, we have found the results are the same for both selection criteria. Therefore, we can use either the AIC or the BIC to determine the true model. The results of our simulation study show the information based criteria are able to distinguish models quite well. Although it has greater difficulty distinguishing between models with dispersion parameters close to each other, it is able to distinguish between the Bernoulli model and COM-Poisson ($\nu = 0.5$) model with a high degree of accuracy. The same conclusion was reached using the likelihood ratio test. Finally, we used our proposed model on a real set of data pertaining to the SARS-CoV-2 pandemic of 2020. Our analysis shows for this particular data set, the geometric, Poisson, and Bernoulli distributions are all viable options for the true competing pathogen distribution.

6.1 Future Works

In this section, we will discuss some future research topics that would be natural extensions of the work conducted in this thesis.

6.2 Other distributions for the competing cause and lifetime variables

In this thesis, we considered the COM-Poisson and Weibull distributions to represent the distribution for the number of competing pathogens and the promotion times of the pathogens, respectively. While our model has clear biological interpretations and demonstrates flexibility, future work may consider the use of other distributions to represent these random variables. Future work may consider the use of the generalized gamma distribution to represent the promotion times to allow for more flexibility. Furthermore, we may consider the generalized power series distribution to model the number of competing pathogens, as done by Borges et al. [14]. The likelihood inference corresponding to these generalized distributions will be of great interest to develop. Furthermore, future work may wish to investigate different distributions to model the number of competing pathogens. Such distributions may include the Gompertz, Yule-Simon, or polyalgorithm distributions.

6.3 Semi-parametric and non-parametric approaches

While the Weibull distribution is a commonly used distribution to represent lifetime in a parametric setup, future works may wish to consider a semi-parametric or non-parametric approach to modeling the promotion time distributions. For example, Balakrishnan et al. [10] used a Cox proportional hazard model with a Weibull baseline hazard function for the promotion times of competing malignant cells related to the occurrence of a tumor. This model reduced to a Weibull distribution with shape parameter γ_0 and scale parameter $\gamma_1 \exp(-\mathbf{x}_c' \boldsymbol{\gamma_2}/\gamma_0)$ where $\mathbf{x}_c = (x_1, \cdots, x_p)'$ is a vector of p covariates and $\boldsymbol{\gamma_2} = (\gamma_{21}, \cdots, \gamma_{2p})'$ is the proportional hazards regression coefficients. By using this model, we can introduce heterogeneity between patients with respect to the pathogen promotion time, which will have great biological significance. Furthermore, we may wish in the future to study a non-parametric framework such as the model proposed by Peng and Dear [44] for a more general form of analysis.

6.4 Other forms of censoring

While the model described in this thesis assumed right censoring for the data, future research may wish to implement other forms of censoring such as interval censoring, which is a more general form of censoring and includes both right and left censoring as special cases. Furthermore, developing the inference under informative censoring will also be of great interest.

6.5 Destructive cure rate model with multiple exposures

As mentioned in Section 1.2, Pal and Majakwara [41] and Majakwara and Pal [37] extended the works of Rodrigues et al. [47] by studying a destructive cure rate model using the COM-Poisson distribution. In a destructive cure rate model, the original number of risk factors undergo a destructive process. This has interesting biological interpretations, especially when discussing infectious diseases, as this model allows us to study the effects of treatment such as antibiotics or other treatments as a destructive process to slow or stop the spread of an infection.

6.6 Other methods of maximum likelihood estimation

While the EM algorithm is a useful tool to find the maximum likelihood estimates of a model, it does have some issues. As mentioned in Chapter 3, the EM algorithm has trouble estimating the dispersion parameter of the COM-Poisson distribution, which requires us to employ a profile likelihood approach to find the estimate for the dispersion parameters. There is a recent research work work by Pal and Roy [42], where the authors have proposed a non-linear conjugate gradient algorithm that can simultaneously estimate all model parameters. It will be of great interest to apply this technique for the M-step of the EM algorithm. Another possibility is to develop a stochastic version of the EM-algorithm, which can be done along the lines of Davies et al. [21].

Work on some of these problem are currently under progress and other will surely be investigated in the future.

APPENDIX A

PROOF OF THEOREM 2.4.1

$$
\begin{aligned}
S_{pop}(y) = {} & P[M_{t_0} = 0, M_{t_1} = 0] \\
& + P[Z_{1,t_0} > y, ..., Z_{M_{t_0},t_0} > y, M_{t_0} \geq 1, M_{t_1} = 0] \\
& + P[Z_{1,t_1} > y, ..., Z_{M_{t_1},t_1} > y, M_{t_1} \geq 1, M_{t_0} = 0] \\
& + P[Z_{1,t_0} > y, ..., Z_{M_{t_0},t_0} > y, Z_{1,t_1} > y, ..., Z_{M_{t_1},t_1} > y, M_{t_0} \geq 1, M_{t_1} \geq 1] \\
= {} & P[M_{t_0} = 0]P[M_{t_1} = 0] \\
& + P[Z_{1,t_0} > y, ..., Z_{M_{t_0},t_0} > y, M_{t_0} \geq 1]P[M_{t_1} = 0] \\
& + P[Z_{1,t_1} > y, ..., Z_{M_{t_1},t_1} > y, M_{t_1} \geq 1]P[M_{t_0} = 0] \\
& + P[Z_{1,t_0} > y, ..., Z_{M_{t_0},t_0} > y, M_{t_0} \geq 1]P[Z_{1,t_1} > y, ..., Z_{M_{t_1},t_1} > y, M_{t_1} \geq 1] \\
= {} & P[M_{t_0} = 0]P[M_{t_1} = 0] \\
& + P[M_{t_1} = 0]\sum_{k=1}^{\infty} P[Z_{1,t_0} > y, ..., Z_{k,t_0} > y]P[M_{t_0} = k] \\
& + P[M_{t_0} = 0]\sum_{j=1}^{\infty} P[Z_{1,t_1} > y, ..., Z_{j,t_1} > y]P[M_{t_1} = j] \\
& + \sum_{k=1}^{\infty} P[Z_{1,t_0} > y, ..., Z_{k,t_0} > y]P[M_{t_0} = k] \sum_{j=1}^{\infty} P[Z_{1,t_1} > y, ..., Z_{j,t_1} > y]P[M_{t_1} = j] \\
= {} & P[M_{t_0} = 0]P[M_{t_1} = 0] \\
& + P[M_{t_1} = 0]\sum_{k=1}^{\infty} \{S_{t_0}(y)\}^k P[M_{t_0} = k] \\
& + P[M_{t_0} = 0]\sum_{j=1}^{\infty} \{S_{t_1}(y)\}^j P[M_{t_1} = j] \\
& + \left\{ \sum_{k=1}^{\infty} \{S_{t_0}(y)\}^k P[M_{t_0} = k] \right\} \left\{ \sum_{j=1}^{\infty} \{S_{t_1}(y)\}^j P[M_{t_1} = j] \right\}.
\end{aligned}
$$

If we let

$$a_k = \{S_{t_0}(y)\}^k P[M_{t_0} = k] \qquad , k = 0, 1, 2, ...$$

and

$$b_j = \{S_{t_1}(y)\}^j P[M_{t_1} = j] \qquad , j = 0, 1, 2, ...,$$

then, we know

$$\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a_k b_j = a_0 b_0 + b_0 \sum_{k=1}^{\infty} a_k + a_0 \sum_{j=1}^{\infty} b_j + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_k b_j.$$

Hence:

$$\begin{aligned}
S_{pop}(y) &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\{S_{t_0}(y)\}^k P[M_{t_0} = k]\{S_{t_1}(y)\}^j P[M_{t_1} = j] \\
&= \left\{ \sum_{k=0}^{\infty}\{S_{t_0}(y)\}^k P[M_{t_0} = k] \right\}\left\{ \sum_{j=0}^{\infty}\{S_{t_1}(y)\}^j P[M_{t_1} = j] \right\} \\
&= \left\{ \sum_{k=0}^{\infty}\{S_{t_0}(y)\}^k \frac{1}{Z(\theta_{t_0}, \nu)} \frac{\{\theta_{t_0}\}^k}{(k!)^v} \right\}\left\{ \sum_{j=0}^{\infty}\{S_{t_1}(y)\}^j \frac{1}{Z(\theta_{t_1}, \nu)} \frac{\{\theta_{t_1}\}^j}{(j!)^v} \right\} \\
&= \frac{1}{Z(\theta_{t_0}, \nu)} \frac{1}{Z(\theta_{t_1}, \nu)} \left\{ \sum_{k=0}^{\infty} \frac{\{\theta_{t_0} S_{t_0}(y)\}^k}{(k!)^v} \right\}\left\{ \sum_{j=0}^{\infty} \frac{\{\theta_{t_1} S_{t_j}(y)\}^j}{(j!)^v} \right\} \\
&= \frac{Z(\theta_{t_0} S_{t_0}(y), \nu) Z(\theta_{t_1} S_{t_1}(y), \nu)}{Z(\theta_{t_0}, \nu) Z(\theta_{t_1}, \nu)}
\end{aligned}$$

as desired.

REFERENCES

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716−723.

[2] Balakrishnan, N., Barui, S. and Milienos, F. (2017). Proportional hazards under Conway–Maxwell-Poisson cure rate model and associated inference. *Statistical Methods in Medical Research* **26**, 2055–2077.

[3] Balakrishnan, N., Koutras, M. V., Milienos, F. S. and Pal, S. (2016). Piecewise linear approximations for cure rate models and associated inferential issues. *Methodology and Computing in Applied Probability* **18**, 937−966.

[4] Balakrishnan, N. and Pal, S. (2012). EM algorithm-based likelihood estimation for some cure rate models. *Journal of Statistical Theory and Practice* **6**, 698−724.

[5] Balakrishnan, N. and Pal, S. (2013). Lognormal lifetimes and likelihood-based inference for flexible cure rate models based on COM-Poisson family. *Computational Statistics & Data Analysis* **67**, 41−67.

[6] Balakrishnan, N. and Pal, S. (2015). Likelihood inference for flexible cure rate models with gamma lifetimes. *Communications in Statistics - Theory and Methods* **44**, 4007−4048.

[7] Balakrishnan, N. and Pal, S. (2015). An EM algorithm for the estimation of parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood and information-based methods. *Computational Statistics* **30**, 151−189.

[8] Balakrishnan, N. and Pal, S. (2016). Expectation maximization-based likelihood inference for flexible cure rate models with Weibull lifetimes. *Statistical Methods in Medical Research* **25**, 1535−1563.

[9] Balakrishnan, N. and Pal, S. (2016). Destructive negative binomial cure rate model and EM-based likelihood inference under Weibull lifetime. *Statistics and Probability Letters* **116**, 429−449.

[10] Berkson, J. and Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* **47**, 501−515.

[11] Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society Series B* **11**, 15−53.

[12] Boatwright, P., Borle, S. and Kadane, J. B. (2003). A model of the joint distribution of purchase quantity and timing. *Journal of the American Statistical Association* **98**, 564−572.

[13] Borges, P., Rodrigues, J., and Balakrishnan, N. (2012). Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data. *Computational Statistics & Data Analysis* **6**, 1703−1713.

[14] Campigotto, F. and Weller, E. (2014). Impact of informative censoring on the Kaplan-Meier estimate of progression-free survival in phase II clinical trials. *Journal of Clinical Oncology* **32**, 3068−3074.

[15] Cancho, V.G., Bandyopadhyaya, D., Louzada, F., and Yiqi, B. (2013). The destructive negative binomial cure rate model with a latent activation scheme. *Statistical Methodology* **13**, 48−68.

[16] Chen, M. -H., Ibrahim, J.G, and Sinha, D. (2008). a new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* **94**, 909−919.

[17] Chequer, P., Hearst, N., Hudes, E.S., Castilho, E., Rutherford, G., Loures, L., and Rodrigues, L. (1992). Determinants of survival in adult Brazilian AIDS patients, 1982-1989. the Brazilian stats AIDS program co-ordinators. *AIDS* **6**, 483−487.

[18] Conway, R. W. and Maxwell, W. L. (1962). Network dispatching by the shortest-operation discipline. *Operations Research* **10**, 51−73.

[19] Cox, D. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman & Hall, London.

[20] Davies, K., Pal, S., and Siddiqua, J. A. (2020). Stochastic EM algorithm for generalized exponential cure rate model and an empirical study.*Journal of Applied Statistics* **47**

[21] Dempster, A.P., Laird, N.M., and Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**, 1−38.

[22] Farewll, V.T. (1982). The use of mixture models fo the analysis of survival data with long-term survivors. *Biometrics* **39**, 1−38

[23] Faucett, C.L., Schenker, N., and Taylor, J.M.G(2004). Survival analysis using auxiliary varibles via multiple imputation, with application to AIDS clinical trial data. *Biometrics* **58**, 37−47.

[24] Gallardo, D. I., Bolfarine, H. and Pedroso-de Lima, A. C. (2016). An EM algorithm for estimating the destructive weighted Poisson cure rate model. *Journal of Statistical Computation and Simulation* **86**, 1497−1515.

[25] Gallardo, D.I., Gomez, Y.M., and Casto, M. (2018). A flexible cure rate model based on the polylogarithm distribution. *Journal of Statisitcal Computation and simulation* **88**, 2137−2149.

[26] Gallardo, D. I., Romeo, J. S. and Meyer, R. (2017). A simplified estimation procedure based on the EM algorithm for the power series cure rate model. *Communications in Statistics - Simulation and Computation* **46**, 6342−6359.

[27] Green, J.K.S., Holman, R.C., and Mahoney, M.A. (1989). Survival analysis of hemophilia associated AIDS cases in the US. *American Public Health Association* **79**, 832−835.

[28] Hagan, H., thiede, H., and Des Jarlais, D.C.(2004). Hepatitis C virus infection among injection drug users: survival analysis of time to seroconversion. *Epidemiology* **15**, 543−549.

[29] Hough, G., Langohr, K., Gomez, G., and Curia, A. (2006). Survival analysis applied to sensory shelf life of foods. *Food Science* **68** 359−362.

[30] Ibrahim, J. G., Chen, M. -H. and Sinha, D. (2001). *Bayesian Survival Analysis.* Springer-Verlag, New York.

[31] Jewell, N.P. and Kalbfeisch, J.D. (1992). Marker models in survival analysis and applications to issues associated with AIDS. In: Jewell N.P., Dietz K., Farewll V.T. (eds) *AIDS Epidemiology* Birkhauser, Boston, MA.

[32] Kim, Y.J. and Jhun, M. (2007). Cure rate model with interval censored data. *Statistics in Medicine* **27**, 3−14.

[33] Klebanov, L. B., Rachev, S. T. and Yakovlev, A. Y. (1993). A stochastic model of radiation carcinogenesis: latent time distributions and their properties. *Mathematical Biosciences* **113**, 51−75.

[34] Kokonendji, C.C., Kiesse, T.S., and Balakrishnan, N. (2009). Semiparametric estimation for count data through weighted distributions. *Mathematical Biosciences* **113**, 51−75.

[35] Li, B., Zhang, H., and He, J. (2019). Some characterizations and properties of COM-Poisson random variables. *Communications in Statistics- Theory and Methods* **6**, 13111−1329.

[36] Majakwara, J. and Pal, S. (2019). On some inferential issues for the destructive COM-Poisson-generalized gamma regression cure rate model. *Communications in Statistics - Simulation and Computation* **48**, 3118−3142.

[37] McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions.* 2nd ed. Hoboken, New Jersey: John Wiley & Sons.

[38] Pal, S. and Balakrishnan, N. (2016). Destructive negative binomial cure rate model and EM-based likelihood inference under Weibull lifetime. *Statistics and Probability Letters* **116**, 9−20.

[39] Pal, S. and Balakrishnan, N. (2017). Likelihood inference for COM-Poisson cure rate model with interval-censored data and Weibull lifetimes. *Statistical Methods in Medical Research* **26**, 2093−2113.

[40] Pal, S., Majakwara, J. and Balakrishnan, N. (2018). An EM algorithm for the destructive COM-Poisson regression cure rate model. *Metrika* **81**, 143−171.

[41] Pal, S. and Roy, S. (2019). A new non-linear conjugate gradient algorithm for destructive cure rate model and a simulation study: illustration with negative binomial competing risks. *arXiv:1905.11379*

[42] Panjer, H. (1987). AIDS: Survival analysis of persons testing HIV+. *Working Paper Series in Actuarial Sciences* **12**, 517−530

[43] Peng, Y. and Dear, K.B.G. (2004). A nonparametric mixture model for cure rate estimation. *Biometrics* **25**, 237−243.

[44] Prentice, R.L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika* **61**, 539−544.

[45] Ravi, M., Wobeser, G.A., Taylor, S.M., and Jackson, M.L. (2004). Naturally acquired feline immunodeficiency virus (FIV) infection in cats from western Canada: Prevalence, disease associations, and survival analysis. *The Canadian Veterinary Journal* **51**, 271−276.

[46] Rodrigues, J., de Castro, M., Balakrishnan, N., and Cancho, V. G. (2011). Destructive weighted Poisson cure rate models. *Lifetime Data Analysis* **17**, 333−346.

[47] Rodrigues, J., de Castro, M., Cancho, V.G. and Balakrishnan, N. (2009). COM-Poisson cure rate survival models and an application to cutaneous melanoma data. *Journal of Statistical Planning and Inference* **139**, 3605−3611.

[48] Santos, M.R., Achcar, J.A., and Martinez, E.Z. (2017). Bayesian and maximum likelihood inference for the defective Gompertz cure rate model with covariates: an application to the cervical carcinoma study. *Science and Nature* **39**, 244−258.

[49] Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461−464.

[50] Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society, Series C* **54**, 127−142.

[51] Stacy, E.W. (1962). A generalization of the gamma distribution. *Annals of Mathematical Statistics* **33**, 1187−1192

[52] Struthers, C.A. and Farewell, V.T. (1989). A mixture model for time to AIDS data with left truncation and an uncertain origin. *Biometrika* **76**, 814−817.

[53] Sy, J.P. and Taylor, J.M.G. (2000). Estimation of a Cox proportional hazards cure model. *Biometrics* **56**, 227-236.

[54] Tournoud, M. and Ecochard, R. (2007). Application of the promotion time cure model with time-changing exposure to the study of HIV/AIDS and other infectious diseases. *Statistics in Medicine* **26**, 1008−1021.

[55] Tournoud, M. and Ecochard, R. (2008). Promotion time models with time-changing exposure and heterogeneity: application to infectious diseases. *Biometrical Journal* **50**, 395−407.

[56] Tucker, S.L. and Taylor, J.M.G. (2009). Improved models for tumor cure. *International Journal of Radiation Biology* **70**, 539−553

[57] Wiangnak, P. and Pal, S. (2018). Gamma lifetimes and associated inference for interval censored cure rate model with COM-Poisson competing cause. *Communications in Statistics - Theory and Methods* **47**, 1491−1509.

[58] Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications.* World Scientific, Singapore.

[59] Yakovlev, A. Y., Tsodikov, A. D. and Bass, L. (1993). A stochastic model of hormesis. *Mathematical Biosciences* **116**, 197−219.

[60] Yin, G. and Ibrahim, J.G. (2005). Cure rate models: a unified approach. *The Canadian Journal of Statistics* **33**, 559−570.

BIOGRAPHICAL STATEMENT

Zachry Engel was born in Pouzzouli, Italy in 1994. He received his B.S. degree in Mathematics from the University of Texas at Arlington, TX, USA, in 2016. He has been a Ph.D. student in the University of Texas at Arlington in Mathematics since August of 2016. His research interests focus on statistical computing, survival analysis, and infectious disease modeling.