Counting processes for correlated binary responses

Forrest W. Crawford* and Daniel Zelterman†

Department of Biostatistics
Yale School of Public Health
New Haven, CT USA

Typeset May 9, 2013

Abstract

We propose a class of continuous-time Markov counting processes for analyzing correlated binary data and establish a correspondence between these models and sums of dependent Bernoulli random variables using a technique called “probabilistic embedding”. Our approach generalizes many previous models for correlated outcomes, admits easily interpretable parameterizations, allows different cluster sizes, incorporates ascertainment bias in a natural way, and dramatically simplifies likelihood-based inference. We demonstrate several new models for dependent outcomes, derive explicit likelihood expressions, and provide algorithms for computing maximum likelihood estimates. We show how to incorporate cluster-specific covariates in a regression setting and apply our method to well-known problems from development toxicology and familial disease epidemiology.

Keywords: Continuous-time Markov process, dependent Bernoulli trials, developmental toxicity, familial disease clustering, probabilistic embedding, teratology.

1 Introduction

The simplest statistical model for a collection of n binary outcomes is the binomial distribution, which assumes that responses are independent and identically distributed (Collett, 2002, page 19). However, many investigations have found that the binomial distribution sometimes gives a poor fit to certain types of data (Greenwood and Yule, 1920; Haseman and Soares, 1976; Altham, 1978). This empirical observation, along with suspicions that the mechanism that generates the outcomes might induce dependencies, has encouraged development of more flexible models that account for correlations in responses. Dependent or correlated binary data arise commonly in studies of developmental toxicology and litter size (Williams, 1975; Kupper and Haseman, 1978; Altham, 1978), familial disease aggregation (Liang et al., 1992; Yu and Zelterman, 2002a), or when ascertainment considerations necessitate a biased approach to sampling (Neuhaus and Segal, 1993; Matthews et al., 2008). Groups of dependent responses are often called “clusters”, and in many applications the response of interest is the number of affected units in a cluster with n members. As an illustrative example, Table 1 shows the observed frequencies of 60 cases of interstitial pulmonary fibrosis (IPF) in the siblings of families with at least one case of chronic obstructive pulmonary disease (COPD). The triangular shape of Table 1 arises because the family/cluster size limits the number of individuals that can be affected. The simplicity of the dataset belies its complexity; the binomial model provides a poor fit and researchers have proposed a wide variety of remedies for the dependency of outcomes within clusters.

There is wide interest in correlated responses. Early work on the problem emphasizes generalizations of the binomial distribution (Bahadur, 1959; Williams, 1975; Kupper and Haseman, 1978; Altham, 1978). Several extensions to logistic regression have been developed for correlated binary outcomes via specification of marginal probabilities using pseudo-likelihood and generalized estimating equations (Zeger and Liang, 1986; Prentice, 1988; Connolly and Liang, 1988; Liang et al., 1992; Zhao and Prentice, 1990). More recently,
| Number of siblings | Number of families | Number of affected siblings |
|-------------------|-------------------|---------------------------|
| n                 | 0 1 2 3 4 5       | 6                         |
| 1                 | 48 36 12          |                           |
| 2                 | 23 15 7 1         |                           |
| 3                 | 17 5 7 3 2        |                           |
| 4                 | 7 3 1 0 0         |                           |
| 6                 | 5 1 0 1           | 1                         |

Table 1: Observed frequencies of interstitial pulmonary fibrosis (IPF) in siblings of patients with chronic obstructive pulmonary disease (COPD), from a dataset analyzed by Liang et al. (1992) and later by Yu and Zelterman (2002a). The binomial distribution provides a poor fit to these data, possibly indicating a household or genetic component to disease risk.

Researchers have sought to model the correlations in the outcomes directly. George and Bowman (1995) introduce expressions for the likelihood of a sum of exchangeable Bernoulli variables via a combinatorial argument using joint probabilities of positive outcomes for subsets of responses. In this context, exchangeability means that the joint probability of all the outcomes in a cluster is invariant to permutation of the responses, a notion we define more formally in Section 2.1. The model of George and Bowman (1995) and Bowman and George (1995) is described as saturated because it allows specification of correlations of all orders. George and Kodell (1996) introduce formal hypothesis testing procedures for this class of models, and Yu and Zelterman (2001, 2002a) describe new parametric dependency models and a framework for exact testing. Yu and Zelterman (2001, 2002a) describe new parametric dependency models and a framework for exact testing. Yu and Zelterman (2002b, 2008) describe novel representations of the models of George and Bowman (1995) and Altham (1978) and use it to derive the beta-binomial distribution and likelihoods for a variety of new parametric distributions, including the family history model and incremental risk models, which we utilize in Section 4.

In this work, we take a very different approach: we show that a family of continuous-time Markov counting processes can be represented as sums of correlated Bernoulli random variables. The likelihoods for these processes have simple expressions, obviating the need for combinatorial arguments usually necessary in other contexts (e.g. Altham (1978)). This correspondence is inspired by the well-known uniform property of arrival times in a Poisson process: conditional on \( n \) arrivals by time \( t \), the arrival times of the \( n \) points are distributed as the order statistics of \( n \) independent draws from the uniform distribution on \((0, t)\) (Karlin and Taylor, 1975, page 126). By augmenting the outcome space with an auxiliary variable – in this case, time – we are able to derive several novel results and dramatically simplify inference for dependent counts. The technique we employ is called “probabilistic embedding” (Blom and Holst, 1991; Blom et al., 1994, page 186). To briefly illustrate the concept, consider scattering \( n \) points on the positive real line according to a Markov point process \( X(t) \) with arrival rates \( \mu_0, \mu_1, \ldots, \mu_{n-1} \) whose values may depend on \( n \) or number of arrivals already observed. This is also known as the “generalized Poisson” or “pure birth” process (Karlin and Taylor, 1975, page 119). The distribution of the sum of the dependent Bernoulli variables is given by the conditional probability of the counting process, \( \Pr(X(s) = r \mid X(0) = 0, X(t) = n) \) for \( 0 < s < t \), which has a convenient closed-form expression. Figure 1 shows a visual representation of the embedding for several different counting models. In Section 4, we give six examples, each corresponding to a mechanistic model for dependency in the binary outcomes, and cast them in the counting process framework. Then in Section 5, we apply our approach to problems in childhood disease epidemiology (Table 1) and developmental toxicology. Appendices provide derivations of likelihood expressions and algorithms for maximum likelihood estimation and regression with covariates.

2 Background

2.1 Sums of exchangeable Bernoulli variables

George and Bowman (1995) describe a likelihood framework for sums of exchangeable Bernoulli random variables that depends on knowledge of joint probabilities of subsets of variables taking positive value.
Consider a sequence of $n$ exchangeable Bernoulli variables $Z_1, \ldots, Z_n$. By exchangeability, we mean that the joint probability of a collection of variables taking certain values is invariant to reordering. More formally,

$$\Pr(Z_1 = z_1, \ldots, Z_n = z_n) = \Pr\left(Z_{\pi(1)} = z_{\pi(1)}, \ldots, Z_{\pi(n)} = z_{\pi(n)}\right)$$

for any permutation $\pi$ of the indices $1, 2, \ldots, n$. Now consider the probability that at least $r$ of the $Z_i$'s are positive. By exchangeability, we can express this as the joint probability that the first $r$ are positive and the remainder are either 1 or 0,

$$\Pr(Z_1 = \cdots = Z_r = 1) = \Pr(Z_1 \cap \cdots \cap Z_r, \ Z_{r+1} \cup \cdots \cup Z_n)$$

$$= \sum_{j=0}^{n-r} (-1)^j \binom{n-r}{j} \Pr(Z_1 + \cdots + Z_r = r, \ Z_{r+1} + \cdots + Z_n = j) \quad (1)$$

using the inclusion-exclusion formula for the probability that at least $j$ of $Z_{r+1}, \ldots, Z_n$ are positive [George and Bowman, 1995]. Now let $\lambda_j = \Pr(Z_{i_1} = Z_{i_2} = \cdots = Z_{i_j} = 1)$ be the joint probability that every $Z_i$ for $i \in I_j$ is positive, where the cardinality of the set $I_j$ is $j$. Then (1) becomes

$$\Pr\left(Z_1 = z_1, \ldots, Z_n = z_n \mid \sum_{i=1}^{n} z_i = r\right) = \sum_{j=0}^{n-r} (-1)^j \binom{n-r}{j} \lambda_{r+j}. \quad (2)$$

Letting $Y_n = \sum_{i=1}^{n} Z_i$, we have

$$\Pr(Y_n = r) = \binom{n}{r} \sum_{j=0}^{n-r} (-1)^j \binom{n-r}{j} \lambda_{r+j}. \quad (3)$$

Note that when $\lambda_j = p^j$, then (3) reduces to the binomial probability,

$$\Pr(Y_n = r) = \binom{n}{r} p^r (1-p)^{n-r}.$$

By specifying the joint probabilities $\lambda_j$ for $j = 0, \ldots, n$, we can describe any dependence model for exchangeable Bernoulli variables. We note that the $\lambda_j$'s are sometimes called “marginal” probabilities [Dang et al., 2009], since they express the joint probability of $j$ successes, summed over all possible outcomes of the remaining $n-j$ variables, as in (1). In statistical settings, a model for the sum of $n$ exchangeable binary outcomes is called “saturated” when all the $\lambda_j$'s are allowed to be nonzero, resulting in a dependency model with correlations of all orders between outcomes [George and Bowman, 1995]. A saturated model of clustered data, where each cluster has the same size $n$, is nonparametric in the sense that it does not rely on a particular parameterization of the $\lambda_j$'s.

We note two major issues with the model of [George and Bowman, 1995] given by (3). First, it is unclear how to interpret the joint probabilities $\lambda_j$ or correlations when analyzing data from clusters of different sizes since the number of unknown parameters for each observation is equal to the cluster size. For example, the saturated model for a cluster of size $n$ depends on $\lambda_1, \ldots, \lambda_n$, and the model for a cluster of size $m > n$ depends on $\lambda_1, \ldots, \lambda_n, \ldots, \lambda_m$, and these probabilities may not be the same for $j = 1, \ldots, n$. Second, it can be difficult to find joint probabilities $\lambda_j$ for $j = 0, \ldots, n$ that result in a well-defined probability mass function. George and Bowman (1995) give monotonicity constraints on the $\lambda_j$’s and Stefanescu and Turnbull (2003) provide diagrams of their feasible regions. Often one must often solve a non-trivial combinatorial problem in order to specify the $\lambda_j$’s (see, e.g., Kuk [2004] Pang and Kuk [2005]. Notably, Yu and Zelterman (2002b) define dependent models using the somewhat more natural conditional probabilities $C_n(k) = \Pr(Z_{n+1} = 1 \mid Z_1 + \cdots + Z_n = k)$. This formulation makes sense when one imagines the dependency in a collection of variables as induced by sequential specification of the probability of the next positive result. Notably, Yu and Zelterman (2002b) derive the beta-binomial distribution using a special specification of $C_n(k)$. In the following section, we outline a method for specifying these conditional probabilities in a more general setting.
2.2 Markov counting processes

Now we consider a continuous-time Markov counting process \( X(t) \) that counts the number of arrivals (or points) before time \( t \). When \( k \) points have arrived, the rate of arrival of the next point is \( \mu_k \). Let \( P_{ir}(t) = \Pr(X(t) = r \mid X(0) = i) \) be the probability that at time \( t \) there have been \( r \) arrivals, given that there were \( i \) already at time 0. This probability obeys the forward Kolmogorov equation

\[
\frac{dP_{ir}(t)}{dt} = \mu_{r-1}P_{i,r-1}(t) - \mu_rP_{ir}(t)
\]

where \( \mu_r > 0 \) is the instantaneous rate of the \( r \) + 1st arrival, given that \( r \) have already arrived (Karlin and Taylor 1975, page 119). This counting model is also known as the “generalized Yule” or “pure birth” process. Note that the Markov structure of the process provides the following useful equivalence:

\[
\Pr(X(t) = n \mid X(s) = k) = \Pr(X(t-s) = n \mid X(0) = k) = P_{kn}(t-s)
\]

where \( s < t \) and \( k \leq n \). The homogeneous Poisson process with rate \( \mu_r = \mu \) is perhaps the best-known counting process, with likelihood \( P_{ir}(t) = (\mu t)^r e^{-\mu t} / r! \) at time \( t \).

For a Markov counting process with arbitrary rates \( \mu_k, k = 0, 1, \ldots \), the transition probability is

\[
P_{ir}(t) = \left( \frac{\prod_{k=i}^{r-1} \mu_k}{\sum_{k=i}^r \prod_{\ell \neq k} (\mu \ell - \mu_k)} \right)^{-1} \exp[-\mu_k t]
\]

for \( 0 \leq i \leq r \leq n \) and \( t > 0 \) and \( \mu_k \neq \mu_r \) for all \( i \) and \( r \) (Renshaw 2011, page 65). For a given set of rates \( \{\mu_k\} \), simpler representations of the likelihood (6) are often available, as we show in Section 4.

A useful consequence of the counting process formulation is that the \( \mu_k \)’s have an easy interpretation. Formally, the transition rates of the process are defined by

\[
\mu_k = \lim_{\Delta t \to 0} \Pr(X(t + \Delta t) = k + 1 \mid X(t) = k) / \Delta t.
\]

We can therefore interpret \( \mu_k \)’s as conditional hazards. For example, in a model for the spread of an infectious disease in a household of size \( n \) where \( k \) are currently infected, \( \mu_k \) is the instantaneous hazard of a new infection in the family, given that the family already has \( k \) cases.

Another benefit of the counting process framework is that it can be modified slightly to deal with more complicated ascertainment schemes. For example, in some observational epidemiological studies, only families with one or more affected children are available for study. We can account for ascertainment bias simply by specifying that family \( i \) has \( m_i \) members already affected at time 0, so \( X(0) = m_i \). Then the relevant likelihood to consider is simply \( \Pr(X(1) = r_1 \mid X(0) = m_i) = P_{m_i,r_1} (1 \mid \theta, m_i) \), where \( m_i \) is the size family \( i \).

In the same way, we can easily account for families of different sizes.

3 The connection

There is a correspondence between the George and Bowman (1995) representation and stochastic counting models of the form (1). In this Section, we illustrate this correspondence and show how to construct a Markov point process \( \{X(\tau) : \tau > 0\} \) conditional on \( X(t) = n \), with arrival rates that depend on the number of arrivals that have already occurred (Blom and Holst 1991, Blom et al. 1994, page 186). The Bernoulli trials are said to be “embedded” in the arrival process in the following way. To each Bernoulli variable \( Z_i \), we associate a time \( t_i \), which is the time of the \( i \)th arrival. If \( t_i < s \), where \( s > 0 \) has been chosen in advance, then \( Z_i = 1 \) and otherwise zero. The following construction of the joint process \( (t_k, Z_k), k = 1, \ldots, n \), illustrates the model. We emphasize that this is a probabilistic construction of equivalence between two random variables, \( X(s) \) and \( Y_n \), and not an inference procedure, which we develop in the Appendix.

3.1 Construction of the equivalence

Let \( X(0) = 0 \) and fix \( s \) and \( t \) such that \( 0 < s < t \). To avoid ambiguity, we must specify that \( X(\tau) \) is right-continuous, so if \( t_k \) is the time of the \( k \)th arrival, then \( X(t_k) = k \). Now let \( W_k \) be the waiting time
of the process in state $k - 1$ before jumping to state $k$ so that $t_j = \sum_{k=1}^j W_k$. Conditional on $X(0) = 0$, $X(t) = n$, and the time $t_{k-1}$ of the $(k-1)$st arrival, the density of the waiting time $W_k$ to the $k$th arrival is given by

$$f_k(w|t_{k-1}) = \mu_{k-1} e^{-\mu_{k-1} w} P_{kn}(t - t_{k-1} - w) \frac{P_{kn}(t - t_{k-1})}{P_{k-1,n}(t - t_{k-1})} \mathbb{1}\{w \leq t - t_{k-1}\}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. The construction proceeds as follows: sample $W_1 \in (0,t)$ from $f_1(w|0)$, set $t_1 = W_1$, and let $Z_1 = \mathbb{1}\{t_1 < s\}$. In step $k$, sample $W_k \in (0,t-t_{k-1})$ from $f_k(w|t_{k-1})$, set $t_k = t_{k-1} + W_k$, and let $Z_k = \mathbb{1}\{t_k < s\}$; terminate the procedure in the $n$th step after finding $Z_n$. This algorithm produces a sequence of Bernoulli random variables $Z_1, \ldots, Z_n$ whose sum $Y_n = \sum_{i=1}^n Z_i$ has the following probability distribution

$$\Pr(Y_n = r) = \Pr(Z_1 = \cdots = Z_r = 1, Z_{r+1} = \cdots = Z_n = 0) = \Pr(t_r < s, t_{r+1} > s | t_n < t) = \Pr(X(s) = r | X(0) = 0, X(t) = n)$$

by construction. This probability is expressed more concisely by

$$\Pr(Y_n = r) = P_{0r}(s)P_{rn}(t-s)/P_{0n}(t).$$

This establishes that the random variable $Y_n$ is distributed as $X(s)|\{X(0) = 0, X(t) = n\}$. We refer to (9) as the endpoint-conditioned likelihood.

Before proceeding, we emphasize that the arrival times $t_i$ in the counting process representation are “dummy” or auxiliary variables whose purpose is to aid in analytic solution of the likelihood for $Y_n$. It is not necessary, for example, to consider $t_i$ to be the actual time of infection of individual $i$ in a familial disease model. The times $t_i$ in which the imaginary arrivals are supposed to occur provides a mechanism for inducing dependency in the binary responses. By exchangeability, the order in which the subjects attained their response (equivalently, the order in which the $Z_i$ are drawn) is irrelevant.

### 3.2 The relationship between $\lambda_k$ and $\mu_k$

The joint success probabilities $\lambda_k$ in the model of [George and Bowman (1995)] can be derived recursively from the counting process transition probabilities, which are functions of the arrival/hazard rates $\mu_k$. First, note that the probability of $n$ successes in $n$ exchangeable Bernoulli trials is given by

$$\Pr(Y_n = n) = \lambda_n = P_{0n}(s)P_{nn}(t-s)/P_{0n}(t)$$

in the counting process model. Likewise, the probability of $n - 1$ successes is given by

$$\Pr(Y_n = n - 1) = n[\lambda_{n-1} - \lambda_n] = P_{0,n-1}(s)P_{n-1,n}(t-s)/P_{0n}(t).$$

Rearranging, we find that

$$\lambda_{n-1} = \frac{1}{n} P_{0,n-1}(s)P_{n-1,n}(t-s)/P_{0n}(t) + \lambda_n,$$

and so on until we reach

$$\Pr(Y_n = 0) = \sum_{j=0}^n (-1)^j \binom{n}{j} \lambda_j = P_{00}(s)P_{0n}(t-s)/P_{0n}(t),$$

recovering each joint probability $\lambda_k$ from the collection of arrival/hazard rates in the counting process representation.

### 3.3 The relationship between $C_n(k)$ and $\mu_k$

There is a closer correspondence between the arrival rates $\mu_k$ and the conditional probabilities described by [Yu and Zelterman (2002b)] who define $C_n(k) = \Pr(Z_{n+1} = 1 | Z_1 + \cdots + Z_n = k)$. Note that (7) defines the
arrival/hazard rate $\mu_k$ as the rate of an arrival in an infinitesimal time increment, given that $k$ arrivals have already happened. Intuitively, $C_n(k)$ is the probability of a transition from $k$ to $k+1$ successes and $\mu_k$ is the rate of this event in the continuous-time counting model. Then $C_n(k)$ can be understood as the probability that at least one more arrival happens before time $s$, given that there have already been $k$ arrivals at time $u < s$ and $X(t) = n$. Marginalizing over this probability with respect to the density $P_{0k}(u)$ of $u$, we find

$$C_n(k) = \int_0^t P_{0k}(u) \Pr(X(s) > k \mid X(u) = k, X(t) = n) \, du$$

$$= \sum_{j=k+1}^n P_{jn}(t-s) \int_0^t \frac{P_{0k}(u)P_{kj}(s-u)}{P_{kn}(t-u)} \, du$$

This provides the desired relationship between the conditional probabilities defined by Yu and Zelterman (2002b) and the arrival/hazard rates $\mu_k$.

### 3.4 Likelihood of the counting process

In statistical inference settings, it is often useful to abandon the sum-of-Bernoullis paradigm and use the raw counting process likelihood instead. This perspective arises in view of three observations. First, it is not necessary to arbitrarily choose the time $t$ at which $X(t) = n$. Suppose we have observed $Y_n = r$ and we set $s = 1$ and assume $t > 1$. The marginal distribution of $X(1) | \{X(0) = 0\}$ is given by $P_{0r}(1)$. Second, when $P_{0r}(s)$ is a probability density on $r \in \{0, \ldots, n\}$ for every $s > 0$, as in Sections 4.5 and 4.6 below), there is no need to normalize it by computing the endpoint-conditioned likelihood. We can simply let $s = 1$ and treat $P_{0r}(s = 1)$ as the relevant likelihood. Third, when even when the counting process likelihood is not a proper density, (i.e. $\sum_{r=0}^n P_{0r}(1) < 1$, as in Sections 4.1, 4.2, 4.3 and 4.4), it may be more natural to normalize it as follows:

$$\Pr(Y_n = r) = P_{0r}(1)/\sum_{k=0}^n P_{0k}(1).$$

Whether one uses the endpoint-conditioned form [0] or the raw counting process likelihood depends on the counting model and the interpretation of $Y_n$.

### 4 Six counting process models

We now describe several models in the counting process framework. Some were originally formulated in the Bernoulli sum paradigm, and others are more natural to derive as counting processes. For each model, we specify the arrival/hazard rates $\mu_k$ for $k = 0, \ldots, n-1$ and give an exact expression for the counting process likelihood. We provide the sum-of-Bernoullis likelihood whenever its form is simple or illuminating. Figure 1 shows an example realization of each of the models with its rate function (the sequence of $\mu_k$’s) superimposed.

#### 4.1 Independence

The simplest model for the sum of Bernoulli outcomes assumes independent and identically distributed $Z_k$, resulting in the binomial distribution. We construct this model explicitly using the constant arrival rates in the Poisson process, $\mu_k = \mu$. The counting process transition probability is the familiar Poisson process likelihood at time $s$

$$P_{0r}(s) = \Pr(X(s) = r \mid X(0) = 0, X(t) = n) = (\mu s)^r e^{-\mu s} / r!.$$ 

Conditioning on $X(t) = n$ means that the times of the $n$ arrival events are distributed as the order statistics of $n$ uniform draws from the interval $(0, t)$ (Karlin and Taylor 1975 page 126; Karlin and Taylor 1999 page 403; Ross 1995 page 66). Fixing $s < t$ allows an interpretation of $X(s)$ as the sum of $n$ independent
Figure 1: Illustration of endpoint-conditioned counting processes developed in Section 4 with cluster size \( n = 20 \). The number of arrivals before time \( s \) is \( Y_n = X(s) \). In each plot, the arrival times are marked by vertical lines and the instantaneous arrival rate is shown in gray. All arrival rates are shown to scale. a. Independence/Poisson process (Section 4.1); b. family history model (Section 4.2); c. Yule model (Section 4.3); d. incremental risk model (Section 4.4); e. susceptible model (Section 4.5); and f. susceptible-infective (SI) model (Section 4.6). In the independence/Poisson model (a), arrival times are uniformly distributed on the interval \((0, t)\) and so \( X(s)|\{X(0) = 0, X(t) = n\} \) is binomially distributed. Under the more complicated models (b-f), the distribution of arrival times is not always uniform, and \( X(s)|\{X(0) = 0, X(t) = n\} \) is distributed as the sum of \( n \) dependent Bernoulli variables.
Bernoulli variables with success probability $s/t$. The full likelihood becomes

$$
\Pr(Y_n = r \mid s, t) = \frac{(\mu s)^k e^{-\mu s}}{k!} \frac{[(\mu(t - s)]^{n-k} e^{-\mu(t-s)}}{(n-k)!} \frac{(\mu t)^n e^{-\mu t}}{n!} \left( \frac{s}{t} \right)^k \left( 1 - \frac{s}{t} \right)^{n-k}.
$$

(10)

Therefore $X(s)|\{ X(0) = 0, X(t) = n \}$ is distributed as Binomial($n, s/t$). From (10), we see that $\lambda_k = (s/t)^k$ in (2). Note that the result does not depend on $\mu$, since only the relationship of $s$ and $t$ matters in the uniform case.

### 4.2 Family history

Yu and Zelterman (2002b) introduce the family history model in which the probability of the first positive outcome is $p$ and each of the remaining $n-1$ outcomes is positive with probability $p'$. Their usage of the term “family history” is consistent with medical terminology describing disease risk: a patient has a family history if any member of their family already has the disease. In the analogous counting process model, $\mu_0 = \alpha$ and $\mu_j = \beta$ for $1 \leq j < n$. The waiting time to the first arrival is exponentially distributed with rate $\alpha$, and subsequent arrivals happen as a Poisson process with rate $\beta$. The counting process likelihood is

$$
L(\alpha, \beta) = P_{0r}(1|\alpha, \beta) = \begin{cases} 
 e^{-\alpha} & r = 0 \\
 \frac{\alpha \beta^{r-1}}{\beta - \alpha} \left[ e^{-\alpha} - e^{-\beta} \sum_{j=0}^{r-1} \frac{[(\beta - \alpha)]^j}{j!} \right] & r \geq 1.
\end{cases}
$$

(11)

A derivation of this likelihood is given in the Supplementary Web Materials. The corresponding endpoint-conditioned likelihood is straightforward to compute, but its functional form is not very illuminating, so we omit it here.

### 4.3 Yule

Consider a self-exciting counting process in which each arrival increases the rate of future arrivals by the same increment. Formally, let $\mu_k = k\beta$ for $k \geq 1$. This model is also known as the Yule or “pure birth” process (Bailey, 1964). Note that $\mu_0 = 0$ so we must begin with $X(0) = 1$. The counting process likelihood is given by the geometric probability

$$
L(\beta) = P_{1r}(1|\beta) = e^{-\beta s} (1 - e^{-\beta s})^{r-1}
$$

(12)

and the endpoint-conditioned likelihood is given by

$$
\Pr(Y_n = r) = \binom{n-1}{r-1} p^{r-1} (1-p)^{n-r}
$$

where $p = e^{-\beta t} (e^{\beta s} - 1)/(1 - e^{-\beta t})$. The binomial form of the endpoint-conditioned likelihood arises whenever the rates $\mu_k$ are a linear function of $k$. A brief derivation of the endpoint-conditioned likelihood is given in the Supplementary Web Materials.

### 4.4 Incremental risk

Yu and Zelterman (2002b) introduce the incremental risk model in which the conditional probability of a success increases with the number of successes already observed:

$$
\Pr(Z_{k+1} = 1|Z_1 + \cdots + Z_k = r) = \logit[\alpha + \beta r] = \frac{e^{\alpha + \beta r}}{1 + e^{\alpha + \beta r}}.
$$
Yu and Zelterman do not derive a closed expression for the likelihood and instead calculate it using a recurrence relation. To derive the counting process likelihood, we treat the arrival rate as the inverse logit transform of the above probability, \( \mu_k = \alpha + k\beta \) for \( k \leq n \). The model is similar to the Yule model above, except for the addition of a constant term \( \alpha \), sometimes called the “immigration” rate. Note that \( \mu_k \) in this model does not depend on cluster/family size \( n \). Here we interpret the parameter \( \alpha \) is the baseline risk, which remains the same regardless of how many arrivals have already happened. The total rate \( \mu_k \) therefore has the interpretation 
\[
\text{rate} = (\text{baseline risk}) + (\text{density-dependent risk}) \times (\# \text{ affected}).
\]

The counting process likelihood is
\[
L(\alpha, \beta) = P_{0r}(1 | \alpha, \beta) = \left( \frac{\alpha/\beta + r - 1}{\alpha/\beta - 1} \right) e^{-\alpha(n-r)}(1 - e^{-\beta})^r
\]

The Supplementary Web Material gives a derivation of the likelihood.

### 4.5 Susceptible

In contrast to the previous models, in the susceptible model the conditional probability of a positive response in one variable depends on \( n \) and decreases when other variables are also positive. Suppose a disease occurs with a constant rate \( \alpha \) in each person who is not already affected. When \( k \) people in a family/cluster of size \( n \) are already affected, there remain \( n-k \) who are still susceptible to the disease, and each of experiences a constant risk \( \alpha \), so the rate of new cases is rate = (incidence) \times (\# susceptible) . In the counting process notation, \( \mu_k = \alpha(n-k) \) for \( 0 \leq j < n \). The counting process likelihood is binomial:
\[
L(\alpha) = P_{0}r(1 | \alpha) = n! \frac{\beta^r}{(n-r)!} \prod_{k=0}^{r} \frac{(-1)^{r-k} \Gamma(\alpha/\beta + r - k)}{k!(r-k)!\Gamma(\alpha/\beta + 1)} e^{-\alpha(n-r)}(1 - e^{-\alpha})^r
\]

The Supplementary Web Material gives a derivation.

### 4.6 Susceptible-infective (SI)

Here we combine the susceptible model with an infectivity rate to form a household-level analogue of the susceptible-infected (SI) epidemic model. Suppose the per-person risk of disease from non-household sources (incidence) is \( \alpha \), and the risk of infection per uninfected person contributed by one infected person is \( \beta \). The \( \alpha \) term can be regarded as the disease risk due to extra-household forces. Britton (1997) calls \( \beta \) the “secondary attack rate” of infections due to other infected persons in the same household unit. The rate of a new affected in a family of size \( n \), given \( k \) already affected is
\[
\text{rate} = (\text{incidence}) \times (\# \text{ susceptible}) + (\text{infectivity}) \times (\# \text{ affected}) \times (\# \text{ susceptible}).
\]

Note that \( \beta \) has units 1/person\(^2\). The above assumptions entail
\[
\mu_k = \alpha(n-k) + \beta k(n-k).
\]

The susceptible model results from \( \beta = 0 \), and a pure contagion/infectivity model is obtained by setting \( \alpha = 0 \). In the infectious disease context, testing whether the outcome (positive disease status) clusters in families is equivalent to asking whether \( \beta \) is nonzero. For example, a finding of \( \beta > 0 \) in a familial disease setting might indicate that there is a genetic or household component to disease risk. The counting process likelihood is
\[
P_{0r}(1) = \frac{n!}{(n-r)!} \beta^r \frac{\Gamma(\alpha/\beta + r)}{\Gamma(\alpha/\beta)} \sum_{k=0}^{r} (-1)^{r-k} \frac{e^{-\alpha(n-k)t}}{k!(r-k)!} \prod_{\ell \neq k} \frac{\Gamma(\alpha/\beta + \ell + k - n)}{\Gamma(\alpha/\beta + 1)}.
\]

This expression is derived in the Supplementary Web Material. This model can be quite useful in statistical settings because the parameterization separates the effect of per-capita incidence (\( \alpha \)) from within-cluster infectivity (\( \beta \)). In regression analyses, it is possible to assess how much of the infectivity is due to cluster-level covariates, as we show below in Section 5.2.
| Model (Section) | Estimate | SE  | AIC   | BIC   | Λ    | χ² df | p      |
|----------------|----------|-----|-------|-------|------|-------|--------|
| Binomial (4.1) | \( p = 0.296 \) | 0.032 | 188.1 | 195.3 | 93.0 | 309.66 | 15 \( < 1 \times 10^{-8} \) |
| Family history (4.2) | \( \alpha = 0.507 \) | 0.081 | 217.7 | 224.9 | 107.8 | 944.56 | 14 \( < 1 \times 10^{-8} \) |
| Incremental risk (4.4) | \( \alpha = 0.491 \) | 0.077 | 213.5 | 220.7 | 105.7 | 198.79 | 14 \( < 1 \times 10^{-8} \) |
| Susceptible (4.5) | \( \alpha = 0.350 \) | 0.045 | 188.1 | 195.3 | 93.0 | 313.86 | 15 \( < 1 \times 10^{-8} \) |
| SI (4.6) | \( \alpha = 0.275 \) | 0.044 | 178.8 | 184.0 | 87.4 | 9.62 | 14 | 0.789 |

Table 2: Results for the IPF dataset. Parameter estimates, standard errors, Akaike information criterion (AIC), Bayesian information criterion (BIC), likelihood ratio statistic \( (\Lambda = -2 \log L) \), \( \chi^2 \) statistic, degrees of freedom (df), and p-values for each model are given. The section in which each model is defined is given in parentheses after the model name. We do not fit the Yule model since it requires \( X(0) > 0 \), which is not appropriate for this dataset.

| Number of siblings | Number of families | Number of affected siblings |
|--------------------|--------------------|---------------------------|
| \( n \)            | 0                  | 1            | 2          | 3          | 4          | 5          | 6          |
| 1 48               | 36.46              | 11.54        |             |             |             |             |             |
| 2 23               | 13.27              | 7.21         | 2.52       |             |             |             |             |
| 3 17               | 7.45               | 5.25         | 3.13       | 1.18       |             |             |             |
| 4 7                | 2.33               | 1.91         | 1.46       | 0.93       | 0.37        |             |             |
| 6 5                | 0.96               | 0.91         | 0.88       | 0.82       | 0.70        | 0.50       | 0.23       |

Table 3: Expected frequencies of interstitial pulmonary fibrosis (IPF) in siblings of patients with chronic obstructive pulmonary disease (COPD), based on the estimated parameters of the SI model presented in Table 2. Pearson’s goodness-of-fit test gives \( \chi^2 = 9.622 \) with df= 14 and \( p = 0.789 \).

5 Applications

In this Section, we illustrate the ease of estimation, goodness of fit, and interpretability of the counting process models for two classic datasets originally analyzed by George and Bowman (1995). In the first, we estimate parameters for the IPF dataset in Table 1. In the second, we fit a four-parameter regression model to data from a developmental toxicity experiment. EM algorithms for maximum likelihood estimation routines are outlined in the Appendix; we use the supplemented EM (SEM) algorithm to compute standard errors for the counting process models (Meng and Rubin, 1991).

5.1 IPF in families with COPD

We fit the traditional binomial model and the family history, incremental risk, susceptible, and SI counting process models to the IPF data in Table 1. The existence of significant familial clustering of IPF cases would indicate a genetic or household environmental component to IPF risk. The results are shown in Table 2. The binomial model estimate and standard error agree with previous authors’ results, and the SI model provides a superior fit to the data. Table 3 shows the expected number of families of size \( n \) with \( k \) members with IPF, based on the fitted SI model. Pearson’s goodness-of-fit test gives \( \chi^2 = 9.622 \) (df= 14, \( p = 0.789 \)) and fails to reject the hypothesis that the IPF data come from the distribution induced by the SI counting process. In the SI model, \( \alpha \) has an interpretation as per-family-member incidence and \( \beta \) is the infectivity per contact between affected and unaffected. Most importantly, \( \beta \) is significantly nonzero, indicating substantial familial clustering of IPF cases, and a possible household or genetic risk component.
Researchers exposed pregnant mice to different doses of the herbicide 2,4,5-trichlorophenoxyacetic acid (2,4,5-T) during gestation and recorded the number of implanted fetuses and the number of fetuses that died, had been resorbed, or had a cleft palate (Holson et al., 1992; Chen and Gaylor, 1992). They observed the number of implanted fetuses, number of “affected” fetuses, and the dose of 2,4,5-T for each mouse in the experiment and are given in Table 1 of George and Bowman (1995). The mice were grouped into six levels, receiving doses of 0, 30, 45, 60, 75, or 90 mg/kg of 2,4,5-T. The responses of litter-mates are correlated because the fetuses gestate in the same mother, and we wish to estimate the dose-dependent toxicity of 2,4,5-T. Let $n_i$ be the number of implanted fetuses (cluster size) in dam $i$, let $d_i$ be the dose, and let $r_i$ be the number of fetuses affected. Then letting $z_i = (1, d_i)$ be the cluster-specific covariate vector for the $i$th observation, we form the SI model as follows. In the $i$th observation, we set $\mu_k = \alpha_i (n_i - k) + \beta_i k (n_i - k)$ for $k = 0, \ldots, n_i$, where $\alpha_i = \exp(\phi_0 + \phi_1 d_i)$ and $\beta_i = \exp(\psi_0 + \psi_1 d_i)$. We use a gradient ascent EM algorithm derived in the Appendix to estimate the parameters and standard errors.

The results of the regression are given in Table 4. The first two lines give estimates and standard errors for the elements of $\phi$ and $\psi$. The next lines give $\alpha$ and $\beta$, stratified by different dose level, where the standard errors were obtained by the delta method. Therefore both exogenous incidence and within-cluster effects are significantly related to the number of affected fetuses in this experiment. The baseline incidence and infectivity are very small in the absence of 2,4,5-T, and the “infectivity” of each affected fetus increases with dose.

### 5.2 Developmental toxicity of an herbicide

Researchers exposed pregnant mice to different doses of the herbicide 2,4,5-trichlorophenoxyacetic acid (2,4,5-T) during gestation and recorded the number of implanted fetuses and the number of fetuses that died, had been resorbed, or had a cleft palate (Holson et al., 1992; Chen and Gaylor, 1992). They observed the number of implanted fetuses, number of “affected” fetuses, and the dose of 2,4,5-T for each mouse in the experiment and are given in Table 1 of George and Bowman (1995). The mice were grouped into six levels, receiving doses of 0, 30, 45, 60, 75, or 90 mg/kg of 2,4,5-T. The responses of litter-mates are correlated because the fetuses gestate in the same mother, and we wish to estimate the dose-dependent toxicity of 2,4,5-T. Let $n_i$ be the number of implanted fetuses (cluster size) in dam $i$, let $d_i$ be the dose, and let $r_i$ be the number of fetuses affected. Then letting $z_i = (1, d_i)$ be the cluster-specific covariate vector for the $i$th observation, we form the SI model as follows. In the $i$th observation, we set $\mu_k = \alpha_i (n_i - k) + \beta_i k (n_i - k)$ for $k = 0, \ldots, n_i$, where $\alpha_i = \exp(\phi_0 + \phi_1 d_i)$ and $\beta_i = \exp(\psi_0 + \psi_1 d_i)$. We use a gradient ascent EM algorithm derived in the Appendix to estimate the parameters and standard errors.

The results of the regression are given in Table 4. The first two lines give estimates and standard errors for the elements of $\phi$ and $\psi$. The next lines give $\alpha$ and $\beta$, stratified by different dose level, where the standard errors were obtained by the delta method. Both $\alpha$ and $\beta$ increase with dose level, and $\beta$ increases much more quickly than $\alpha$. Therefore both exogenous incidence and within-cluster effects are significantly related to the number of affected fetuses in this experiment. The baseline incidence and infectivity are very small in the absence of 2,4,5-T, and the “infectivity” of each affected fetus increases with dose.

### 6 Discussion

Even after years of methodological progress, statistical inference for dependent binary responses remains difficult, and the choice of dependency models is often limited by analytic tractability. Unfortunately, models whose parameters have straightforward real-life interpretations (e.g. stochastic epidemic models) can be difficult to fit. Conversely, when a model is chosen mostly for computational convenience (e.g. binomial mixtures or marginal models), it is often unclear how to interpret the estimated parameters in a way that gives insight into the real-life nature of dependency between responses. The paradigm of George and Bowman (1995) is useful because the likelihood for any dependency model can be expressed simply. However, it can be difficult to translate knowledge of the dependency pattern into the joint outcome probabilities necessary to write the likelihood. There has also been uncertainty about how to incorporate data from different cluster
sizes; joint probabilities and higher-order correlations are not the same in general when the number of Bernoulli trials changes (Stefanescu and Turnbull, 2003; Matthews et al., 2005). In addition, ascertainment issues can dramatically complicate inference under this framework. Often clusters are identified through one or more probands satisfying certain criteria related to their response, creating a possible bias in the sample (Wickramaratne, 2004; Matthews et al., 2008).

In this work, we have developed a flexible class of Markov-dependent stochastic models for analyzing clustered binary data. Arrival process rates are often extremely easy to specify; usually a consideration of the conditional risk of positive outcome, given cluster size \(n\), and \(k\) already positive, is enough to express the \(\mu_k\)'s in a useful form. Once this is done, inference is straightforward either by numerical maximization of the likelihood or the EM algorithm approach for the marginal likelihoods outlined in the Appendix. The extension to regression with cluster-specific covariates does not substantially increase the complexity of the modeling or inference tasks. We have established the correspondence between sums of dependent Bernoulli variables and endpoint-conditioned counting processes. However, it is not always necessary to interpret dependent counts as sums of dependent Bernoulli variables. Counting processes can be easily defined so that the arrival rates are functions of cluster size \(n\), thereby accounting for clusters of different sizes. Any endpoint-conditioned counting process can be expressed as a sum of dependent Bernoulli variables, and this provides theoretical and philosophical justification for its use, but is not necessary to address the statistical inference task, as we have shown in our examples. We do not claim that every dependence model for \(n\) exchangeable Bernoulli variables can be represented as an endpoint-conditioned Markov counting process. Our argument is the converse: a counting process representation can be used to derive a variety of interesting continuous-time models, and this representation can always be interpreted as a sum of Bernoulli variables with certain dependent success probabilities.

We wish to make a special note of our formulation of the susceptible-infectious model in section 4.6. Of all the counting process models we present in this work, we believe the SI model is most useful. Statistical inference under this model addresses a fundamental question in infectious disease and genetic epidemiology: estimating \(\beta > 0\) means that some disease risk is due to genetic or household effects. The functional form of the rates in the SI model mirror those in the formulation of stochastic SIR type models (Andersson and Britton, 2000), and the units of the parameters are sensible and familiar to epidemiologists: the incidence rate \(\alpha\) has units 1/person and the contact/infectivity rate \(\beta\) has units 1/person\(^2\).

References

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statistics* **27**, 162–167.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and their Statistical Analysis*. Lecture notes in statistics. Springer New York.

Bahadur, R. (1959). A representation of the joint distribution of responses to \(n\) dichotomous items. Technical report, DTIC Document.

Bailey, N. T. J. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley New York.

Blom, G. and Holst, L. (1991). Embedding procedures for discrete problems in probability. *Mathematical Scientist* **16**, 29–40.

Blom, G., Holst, L., and Sandell, D. (1994). *Problems and Snapshots from the World of Probability*. Springer.

Bowman, D. and George, E. (1995). A saturated model for analyzing exchangeable binary data: Applications to clinical and developmental toxicity studies. *Journal of the American Statistical Association* **90**, 871–879.

Britton, T. (1997). Tests to detect clustering of infected individuals within families. *Biometrics* **53**, 98–109.

Chen, J. J. and Gaylor, D. W. (1992). Correlations of developmental end points observed after 2,4,5-trichlorophenoxyacetic acid exposure in mice. *Teratology* **45**, 241–246.

Collett, D. (2002). *Modelling binary data*. Chapman & Hall/CRC.
Connolly, M. and Liang, K. (1988). Conditional logistic regression models for correlated binary data. *Biometrika* **75**, 501–506.

Crawford, F. W., Minin, V. N., and Suchard, M. A. (2012). Estimation for general birth-death processes. *arXiv preprint arXiv:1111.4954*.

Dang, X., Keeton, S., and Peng, H. (2009). A unified approach for analyzing exchangeable binary data with applications to developmental toxicity studies. *Statistics in Medicine* **28**, 2580–2604.

De Finetti, B. (1931). *Funzione caratteristica di un fenomeno aleatorio*. Academia Nazionale del Linceo.

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 B* **39**, 1–38.

Draper, D., Hodges, J., Mallows, C., and Pregibon, D. (1993). Exchangeability and data analysis. *Journal of the Royal Statistical Society A* **156**, 9–37.

George, E. and Bowman, D. (1995). A full likelihood procedure for analysing exchangeable binary data. *Biometrics* **51**, 512–523.

George, E. and Kodell, R. (1996). Tests of independence, treatment heterogeneity dose-related trend with exchangeable binary data. *Journal of the American Statistical Association* **91**, 1602–1610.

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* **83**, 255–279.

Haseman, J. and Soares, E. (1976). The distribution of fetal death control mice and its implications on statistical tests for dominant lethal effects. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **41**, 277–287.

Holson, J., Gaines, T., Nelson, C., LaBorde, J., Gaylor, D., Sheehan, D., and Young, J. (1992). Developmental toxicity of 2,4,5-trichlorophenoxyacetic acid (2,4,5-t): I. multireplicated dose-response studies in four inbred strains and one outbred stock of mice. *Fundamental and Applied Toxicology* **19**, 286–297.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*. Academic Press.

Karlin, S. and Taylor, H. M. (1999). *A Second Course in Stochastic Processes*. Academic Press.

Kuk, A. Y. (2004). A litter-based approach to risk assessment in developmental toxicity studies via a power family of completely monotone functions. *Journal of the Royal Statistical Society C* **53**, 369–386.

Kupper, L. and Haseman, J. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* pages 69–76.

Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer New York, 2nd edition.

Liang, K., Zeger, S., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society B* **54**, 3–40.

Matthews, A., Finkelstein, D., and Betensky, R. (2005). Analysis of familial aggregation in the presence of varying family sizes. *Journal of the Royal Statistical Society C* **54**, 847–862.

Matthews, A., Finkelstein, D., and Betensky, R. (2008). Analysis of familial aggregation studies with complex ascertainment schemes. *Statistics in Medicine* **27**, 5076–5092.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.

Neuhaus, J. M. and Segal, M. R. (1993). Design effects for binary regression models fitted to dependent data. *Statistics in Medicine* **12**, 1259–1268.
A.1 Maximum likelihood estimation

Let \( \theta \) be a vector of parameters that control the arrival rates \( \mu_k \) for \( k = 0, \ldots, n - 1 \); we write \( \mu_k(\theta) \) for each \( k \). For example, in (14) of Section 4.6, \( \theta = (\alpha, \beta)' \). Let \( r_i \) be the number of positive responses among the \( n_i \) units in cluster \( i \) and the counting process likelihood is the product of the cluster likelihoods. We can often compute the likelihood directly, but it can be difficult to maximize analytically. The arrival process interpretation provides an alternative perspective via maximum likelihood inference using the EM algorithm (Dempster et al., 1977). Suppose we did observe the arrival times of the \( r \) points, \( t_1, \ldots, t_r \) for a single cluster. If \( X(0) = 0 \) and \( X(1) = r \), then the log-likelihood of that cluster becomes

\[
\ell(\theta) = \sum_{k=0}^{r} 1\{k < r\} \log \mu_k(\theta) - \mu_k(\theta)(t_k - t_{k-1})
\]  

(15)

The problem with (15) is that we have not observed the waiting times between arrivals. We therefore consider the random variable \( W_k = t_k - t_{k-1} \), let \( \hat{\theta}(m) \) be a current parameter estimate, and form the surrogate log-likelihood function

\[
Q \left( \theta \mid \hat{\theta}(m) \right) = \sum_{k=0}^{r} 1\{k < r\} \log \mu_k(\theta) - \mu_k(\theta)\mathbb{E} \left[ W_k \mid r, \hat{\theta}(m) \right]
\]  

(16)
where the expectation is conditional on \( r \) arrivals and a previous parameter estimate. By maximizing \( Q \) to find \( \theta^{(m+1)} \) and re-forming \( Q \), we have an EM algorithm. We obtain asymptotic standard errors using the SEM algorithm to compute the information matrix after the MLE has been found (Meng and Rubin 1991). The conditional expectation in (10) is the expected waiting time between the \((k-1)\)st and the \(k\)th arrival, given the starting state \( x_0 \) and ending state \( n \) of the process, given by

\[
\mathbb{E}(W_k | Y) = \int_0^t P_{x_0 k}(s) P_{kn}(1-s) \, ds / P_{x0n}(1).
\]

### A.1.1 Inference for the SI model

Our presentation here follows that of Crawford et al. (2012), who develop EM algorithms for general birth-death processes. The unknown parameter vector is \( \theta = (\alpha, \beta)' \). The surrogate log-likelihood \( Q \) becomes

\[
Q(\theta | \theta^{(m)}) = \sum_{i=1}^N \left[ \frac{r_i - 1}{k=0} \sum_{k=0}^r \log(\alpha + \beta k) - \sum_{k=0}^r \left( \alpha(n_i - k) + \beta k(n_i - k) \right) \mathbb{E}[W_k | r_i, \theta^{(m)}] \right] (17)
\]

but this function is difficult to maximize. However, we can take advantage of a minorizing inequality,

\[
\log(\alpha + \beta k) \leq p_k \log(p_k \alpha) + (1 - p_k) \log((1 - p_k) / \beta k)
\]

where \( p_k = \alpha^{(m)} / (\alpha^{(m)} + k \beta^{(m)}) \). Now define a new surrogate function as follows:

\[
H(\theta | \theta^{(m)}) = \sum_{i=1}^N \left[ \frac{r_i - 1}{k=0} \sum_{k=0}^r p_k \log(p_k \alpha) + (1 - p_k) \log((1 - p_k) / \beta k) - \sum_{k=0}^r \left( \alpha(n_i - k) + \beta k(n_i - k) \right) \mathbb{E}[W_k | r_i] \right]. (18)
\]

Now we have \( H(\theta | \theta^{(m)}) \leq Q(\theta | \theta^{(m)}) \) for all \( \theta \) with equality when \( \theta = \theta^{(m)} = (\alpha^{(m)}, \beta^{(m)}) \). This results in an EM-MM (Minorize-Maximize) algorithm; the ascent property of the original EM algorithm is preserved (Lange 2010). Differentiating \( H \) and solving for the unknown parameters, the updates are then

\[
\alpha^{(m+1)} = \frac{\sum_{i=1}^N \sum_{k=0}^{r_i - 1} p_k (n_i - k) \mathbb{E}[W_k | r_i]}{\sum_{i=1}^N \sum_{k=0}^{r_i - 1} (1 - p_k)} \quad \text{and} \quad \beta^{(m+1)} = \frac{\sum_{i=1}^N \sum_{k=0}^{r_i - 1} (1 - p_k)}{\sum_{i=1}^N \sum_{k=0}^{r_i - 1} k(n_i - k) \mathbb{E}[W_k | r_i]}.
\]

### A.1.2 Inference for the SI regression model

Now consider the SI model above with incidence and infectivity parameters that are themselves functions of cluster-specific covariates. Let \( \alpha = \exp(z_i' \phi) \) and \( \beta = \exp(x_i' \psi) \), where \( z_i \) and \( x_i \) are vectors of covariates for cluster \( i \), and \( \phi \) and \( \psi \) are unknown parameter vectors of corresponding dimension. We seek to estimate \( \theta = (\phi, \psi)' \). The surrogate becomes

\[
H(\theta) = \sum_{i=1}^N \sum_{k=0}^{r_i - 1} p_k z_i' \phi + (1 - p_k) x_i' \psi - \sum_{k=0}^{r_i} \left( \exp[z_i' \phi] (n_i - k) + \exp[x_i' \psi] k(n_i - k) \right) \mathbb{E}[W_k | r_i].
\]

A Newton-Raphson algorithm follows via the update \( \theta^{(m+1)} = \theta^{(m)} - (d^2 H)^{-1} \nabla_\theta H \).
Supplementary Materials

Derivation of the likelihood for the family history model

In the family history model, the first arrival happens at an exponentially distributed time with rate $\alpha$. Subsequent arrivals happen as a Poisson process with rate $\beta$. The probability that no arrivals happen before time $t$ is the probability that the first arrival happens after time $t$ is $P_{00}(t) = e^{-\alpha t}$. For $r > 0$, we condition on the time $u$ of the first arrival, which happens with probability $\alpha e^{-\alpha u}$ and integrate it out with respect to the probability of $r - 1$ arrivals in the remaining time as follows:

$$P_{0r}(t) = \int_0^t \alpha e^{-\alpha u} \frac{[\beta(t-u)]^{r-1} e^{-\beta(t-u)}}{(r-1)!} du$$

$$= \frac{\alpha \beta^{r-1}}{(r-1)!} \int_0^t e^{-(\alpha-\beta)u(t-u)} du$$

$$= \frac{\alpha \beta^{r-1}}{(\beta-\alpha)^r} \left[ e^{-\alpha t} - e^{-\beta t} \sum_{j=0}^{r-1} \frac{[t(\beta-\alpha)]^j}{j!} \right].$$

Derivation of the likelihood for the Yule model

In the Yule process, $\mu_k = k\beta$. The well-known counting process likelihood is derived by Bailey (1964). The endpoint-conditioned likelihood is given by

$$\frac{P_{1r}(s)P_{rn}(t-s)}{P_{1n}} = \frac{e^{-\beta t}(1-e^{-\beta s})^{r-1}e^{-\beta(t-s)}(1-e^{-\beta(t-s)})^{n-r}}{e^{-\beta t}(1-e^{-\beta t})^n}$$

$$= \binom{n-1}{r-1} \left[ e^{-\beta(t-s)} - e^{-\beta t} \right]^{r-1} \left[ 1 - e^{-\beta(t-s)} \right]^{n-r}$$

$$= \binom{n-1}{r-1} p^{r-1}(1-p)^{n-r}$$

where $p = e^{-\beta t}(e^{\beta s} - 1)/(1 - e^{-\beta t})$. Note that we have conditioned on $X(0) = 1$ because $\mu_0 = 0$ in the Yule process.

Derivation of the likelihood for the incremental risk model

Using $\mu_k = \alpha + k\beta$, we find

$$P_{ij}(t) = \prod_{k=i}^{i-1}(\alpha + k\beta) \sum_{k=i}^{j} \frac{e^{-(\alpha+k\beta)t}}{\prod_{\ell\neq k}\beta(\ell-k)}$$

$$= \frac{\Gamma(\alpha/\beta + j)}{\Gamma(\alpha/\beta + i)} e^{-\alpha t} \sum_{k=i}^{j} (-1)^k \frac{e^{-k\beta t}}{(k-i)!(j-k)!}$$

$$= \frac{\Gamma(\alpha/\beta + j)!}{\Gamma(\alpha/\beta + i)!} e^{-\alpha t} \sum_{k=i}^{j} (-1)^{k-i} \binom{k}{i} \binom{j}{k} e^{-k\beta t}. $$

Note that setting $i = 0$ and $j = r$ gives

$$P_{0r}(t) = \left(\frac{\alpha/\beta + r - 1}{\alpha/\beta - 1}\right) e^{-\alpha t} (1-e^{-\beta t})^r,$$
where
\[
\left( \frac{\alpha/\beta + r - 1}{\alpha/\beta - 1} \right) = \frac{\Gamma(\alpha/\beta + r)}{\Gamma(\alpha/\beta) r!}.
\]
The endpoint-conditioned likelihood is
\[
\frac{P_{0r}(s)P_{rn}(t-s)}{P_{0n}(t)} = \binom{n}{r} p^r (1-p)^{n-r}
\]
where \(p = (e^{-\beta s} - e^{-\beta t})/(1 - e^{-\beta t})\).

**Derivation of the likelihood for the susceptible model**

Using \(\mu_k = \alpha(n - k)\), we have the binomial likelihood
\[
P_{ij}(t) = \prod_{k=i}^{j-1} (n-k) \sum_{k=i}^{j} \frac{e^{-\alpha(n-k)t}}{\prod_{\ell\neq k}(k-\ell)}
\]
\[
= \binom{n-i}{n-j} \sum_{k=i}^{j} \frac{(-1)^{j-k} e^{-\alpha(n-k)t}}{(k-i)! (j-k)!}
\]
\[
= \binom{n-i}{j-i} e^{-\alpha(n-j)} (1 - e^{-\alpha t})^{j-i}.
\]
and the endpoint-conditioned likelihood again has binomial form
\[
\frac{P_{0r}(s)P_{rn}(t-s)}{P_{0n}(t)} = \binom{n}{r} p^r (1-p)^{n-r}
\]
where \(p = (e^{-\alpha s} - e^{-\alpha t})/(1 - e^{-\alpha t})\).

**Derivation of the likelihood for the SI model**

Using \(\mu_k = \alpha(n - k) + \beta k(n - k)\), we have
\[
P_{0r}(t) = \prod_{k=0}^{r-1} (n-k)(\alpha + \beta k) \sum_{k=0}^{r} \frac{e^{-\alpha(n-k)t}}{\prod_{\ell\neq k}(\ell-k)\beta(n-\ell-k) - \alpha}
\]
\[
= \frac{n!}{(n-r)!} \beta^r \frac{\Gamma(\alpha/\beta + r)}{\Gamma(\alpha/\beta)} \sum_{k=0}^{r} (-1)^{r-k} \frac{e^{-\alpha(n-k)t}}{k!(r-k)! \prod_{\ell\neq k}[\alpha + \beta(\ell + k - n)]}.
\]