J. R. Statist. Soc. B (1996)
58, *No.* 4, *pp.* 751–762

Counting Process Models for Infectious Disease Data: Distinguishing Exposure to Infection from Susceptibility

By PHILIP H. RHODES, M. ELIZABETH HALLORAN† and IRA M. LONGINI, JR

Centers for Disease Control, Atlanta, USA *Emory University, Atlanta, USA*

[Received April 1995. Final revision December 1995]

SUMMARY

Differences in infection rates among types of individuals within a population can arise from differences in amount of exposure to infection or from differences in susceptibility to infection. We derive models for infection rates that incorporate contact rates between individuals and variables affecting susceptibility to infection. We emphasize the distinction between controlling for exposure opportunity (expected exposure) and actual exposure. We present a marked counting process model for the combined contact and infection transmission processes. When the contact process is not observable, we develop thinned counting process models that reduce to a proportional hazards model. We show that the different commonly used parameters for evaluating covariate effects, such as vaccine efficacy, form a hierarchy depending on the amount of information available about the components of the transmission system.

Keywords: COUNTING PROCESSES; INFECTIOUS DISEASE MODELS; MARKED COUNTING PROCESS; THINNED COUNTING PROCESS; VACCINE EFFICACY

1. INTRODUCTION

Differences in infectious disease rates among types of individuals can arise either from differences in the exposure to infection or from differences in the susceptibility. Several common parameters of relative risk are used to estimate the effect of covariates, such as genetic factors, vaccination status, chemoprophylaxis or age, on susceptibility. For example, relative transmission probabilities (secondary attack rates), relative person–time measures, hazard ratios and relative cumulative incidence (attack rates) have all historically been used to estimate vaccine efficacy. Quite commonly two or more methods of analysis are presented in the same paper.

Until now, however, there has been little effort to relate the different measures to one another formally, or their interpretation in terms of the underlying contact and infection processes. Here we extend counting process models for infection rates (Becker, 1982, 1985, 1989) to incorporate contact rates between individuals, infectiousness of the infectives and variables affecting susceptibility to infection given that such a contact has occurred. Using these counting process models, we demonstrate that the commonly used relative risk parameters form a hierarchy requiring different amounts of information about the contact and infection process. We emphasize the distinction between exposure opportunity and actual exposure, and the amount of information that we have about these. Separation of the contact and infection

†*Address for correspondence*: Department of Biostatistics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA.
E-mail: betz@bear.sph.emory.edu

 0035–9246/96/58751

process allows quantification of the different contributions of the contact process, infectiousness and susceptibility in the estimated relative risk of infection in the comparison groups. We provide guidelines for choosing between study designs requiring different amounts of information.

2. CONTACT, INFECTION, SUSCEPTIBILITY AND INFECTIOUSNESS PROCESSES

2.1. *Overview of Hierarchy of Information Levels*

Table 1 contains an overview of the hierarchy of levels of information that could be known about a population of interacting individual hosts with a circulating infectious agent in it. At a minimum, we need to know those covariates that are relevant to susceptibility as well as who is actually susceptible. The hierarchy goes from level I to IV, or from (a) to (f), as information is either lost or ignored. In (a), we know all contacts between individuals, whereas in (b) we only know when infective individuals contact susceptibles. Level (b) is analogous to a vaccine efficacy study using the household secondary attack rate, studies in tuberculosis using contact tracing to estimate transmission probabilities or discordant partner studies to estimate the transmission probability of human immunodeficiency virus. Levels IIA and IIB, or (c) and (d), have information only on contacts that lead to infection, or the times at which individuals are infectious respectively. These levels have important differences, but share enough similarities that they are developed in tandem. The analysis of the former has the form of a Poisson regression. At level III, we know just the infection times, which under certain conditions leads to a stratified Cox regression analysis. Finally, at level IV, we only know that a person becomes infected sometime during the study period. This provides information for an analysis based on cumulative incidence or the distribution function, such as vaccine efficacy based on attack rates. In the next sections we present the formal counting process models for these differing histories, demonstrating the hierarchy of the parameters.

2.2. *Notation and Definitions*

All processes defined below occur in continuous time and are orderly, i.e. multiple points do not occur at any time t. Also, there are no tied jumps for pairs of processes of the same type involving different individuals; for example, no two infections can

TABLE 1
Levels and amount of information for each history

Level		*Type of information for each history*
I	(a)	All contacts between individuals and outcomes of those contacts (whether an infection is transmitted)
	(b)	Only those contacts between infective and susceptible individuals and infection outcome of those contacts
IIA	(c)	Only contacts leading to infections (who infects whom)
IIB	(d)	Infectious periods, i.e. the times at which individuals become and cease to be infectious
III	(e)	The times at which individuals become infected
IV	(f)	Whether or not an infection occurs to each individual in some time period $(0, T]$

occur at the same time. Some pairs of processes of different types may jump at the same time (for example, see C_{ij} and N_{ij} below). Consider a closed population of n individuals. Let $C_{ij}(t)$ be the counting process for person j contacting person i ($j \to i$), $i, j = 1, \ldots, n$, $i \neq j$. We set $C_{ij}(0) = 0$ for all i, j, i.e. we disregard all contacts that occur before the start of the study. For a study of length T, let t_{ijk} represent times in $(0, T]$ at which $j \to i$, $k = 1, \ldots, C_{ij}(T) = c_{ij}$. For an epidemic, T refers either to the end of the epidemic or to some preset ending time. For an endemic situation, T is some selected time at which an analysis is to be performed. Technically, we require that T be a stopping time with respect to some appropriate history (Bremaud, 1981).

Let $N_{ij}(t)$ be the counting process for the process j infects i, i.e. $dN_{ij}(t) = 1$ if person j infects person i at time t. Let δ_{ijk} be an indicator variable for whether the contact at t_{ijk} results in an infection (i.e. $\delta_{ijk} = dN_{ij}(t_{ijk})$). Let $N_{i.}(t) = \Sigma_j N_{ij}(t)$. Let $\delta_i = N_{i.}(T) - N_{i.}(0)$, i.e. $\delta_i = 1$ if person i becomes infected in $(0, T]$ and $\delta_i = 0$ if not. Alternatively, we may view the process N_{ij} as a random variable associated with each jump of C_{ij} that indicated whether or not an infection was transmitted (Bremaud, 1981). It is possible that $N_{i.}(0) = 1$ which indicates that person i was infected before the start of the current study. However, for the analyses considered here, we are interested only in counting infections that occur after time 0. We assume that the infection can occur at most once, i.e. $N_{i.}(t) \leqslant 1$.

Let $I_j(t) = 1$ if person j is infectious at time t and $I_j(t) = 0$ otherwise. A person is infectious immediately after becoming infected (no latent period). Let $S_i(t) = 1$ if person i is susceptible at time t and $S_i(t) = 0$ otherwise. We define both sets of these processes to be left continuous. Thus, I_j and S_i are predictable processes (Bremaud, 1981).

2.3. *Intensities for Contact Processes*

Let the intensity of the contact process C_{ij} be denoted by $\lambda_{ij}(t)$ ($\lambda_{ii}(t) = 0$), i.e.

$$\lambda_{ij}(t) = \lim_{\Delta \to 0} \left(\frac{\Pr[\{C_{ij}(t + \Delta) - C_{ij}(t)\} = 1 | \mathcal{H}_t]}{\Delta} \right), \tag{1}$$

where $\mathcal{H}_t$ is some history (Bremaud, 1981). Informally, by a history we mean some observed information arising from various processes on the time interval $(0, t]$. Technically, $\mathcal{H}_t$ is a σ-algebra generated by these processes on $(0, t]$. There may be several such histories that are of interest. We shall assume that the λ_{ij} are constants that can be parameterized by using covariates $\mathbf{G}_i$ and $\mathbf{G}_j$ and a set of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_R)$, where $R \ll n(n-1)$, the number of pairs of individuals.

More generally, the contact rates could vary over time, such as cyclically, or be history dependent. For example, the occurrence of an infection could cause a person j to reduce his or her activity and thus to lower the intensities λ_{ij} for all i. We do not consider this aspect further, and we drop the notation for $\mathbf{G}_j$.

2.4. *Intensities for Infection Processes*

Consider any C_{ij} contact process discussed earlier. The contact process plus the infection outcomes, δ_{ijk}, constitute a marked counting process (Bremaud, 1981; Arjas, 1989). Consider the multivariate infection process $\mathbf{N}(t) = \{N_{1.}(t), \ldots, N_{n.}(t)\}$. The process $N_{..}(t) = \Sigma_{i=1}^{n} N_{i.}(t)$ plus the identity and covariate values of the person

infected at each jump are also a marked counting process. Let the function $\rho(t)$ denote the probability that an event occurring at time t in the original process will be retained by a thinned process. If $\lambda(t)$ is an intensity for the original process and $\rho(t)$ is predictable, the intensity for the thinned process is $\rho(t)\,\lambda(t)$ (Bremaud, 1981).

Each infection process N_{ij} is a thinned version of the corresponding contact process C_{ij}. Let $p(t; \mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\beta})$ represent the probability that a contact $j \to i$ at time t results in an infection if person j is infectious and person i is susceptible. This is also called the transmission probability. The $\mathbf{z}_i$ are covariates associated with susceptible i, and $\mathbf{z}_j$ covariates associated with infective j, whereas $\boldsymbol{\beta}$ is a vector of unknown parameters. If either $I_j(t)$ or $S_i(t)$ is 0, a point from C_{ij} has probability 0 of being accepted. If both $I_j(t)$ and $S_i(t)$ are 1, the point is accepted with probability

$$p(t; \mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\beta})\, S_i(t)\, I_j(t).$$

The time- and history-dependent probability $\rho_{ij}(t)$ that a point from C_{ij} will be accepted for N_{ij} is $p(t; \mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\beta})$. A dependence on $\mathbf{z}_j$ implies that individuals are differentially infectious. We assume in this paper that all infectives are equally infectious, and we drop the dependence on $\mathbf{z}_j$. An intensity for $N_{ij}(t)$ may then be written as

$$\alpha_{ij}(t) = \lambda_{ij}(t)\, p(t; \mathbf{z}_i, \boldsymbol{\beta})\, S_i(t)\, I_j(t), \tag{2}$$

where the infection process is a thinned version of the contact process.

3. INFORMATION LEVELS AND TYPES OF STATISTICAL ANALYSIS

In this section, we derive an appropriate statistical analysis for the transmission parameters based on the properties of marked or thinned counting processes. $\mathbf{Z}_i$ and $\mathbf{G}_i$ denote covariates associated with the susceptibility and contact parameters respectively. In most of the development here, the covariates associated with the contact parameters are assumed to be the same for all individuals.

3.1. *Level I*

In the first level of information, either all contacts between individuals and outcomes of those contacts are known, or contacts between infectives and the susceptibles whom they contact during their infectious period:

$$\mathcal{H}_t^{\mathrm{I}} = \sigma\{C_{ij}(s), N_{ij}(t), I_j(s), S_i(s), \mathbf{Z}_i(s), \mathbf{G}_i(s), 0 \leqslant s \leqslant t\}.$$

The analysis remains the same for evaluating covariates related to susceptibility since only contacts between infectives and susceptibles enter the analysis. Estimation of the contact process will differ, however. The log-likelihood of observing contacts at the set of points $\{t_{ijk}: i, j = 1, \ldots, n, k = 1, \ldots, C_{ij}(T)\}$ (Fleming and Harrington, 1991) is given below in terms of stochastic integrals:

$$\log L(C) = \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^T \log\, \lambda_{ij}(t)\, \mathrm{d}C_{ij}(t) - \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^T \lambda_{ij}(t)\, \mathrm{d}t. \tag{3}$$

Without loss of generality, in the equations that follow, we suppress the time

dependence of the **Z**-covariates. The conditional likelihood for the infection outcome marks (the N_{ij}-processes) given the C_{ij}-, $\mathbf{Z}_i$-, S_i- and I_j-processes is

$$\prod_{i=1}^{n}\prod_{j=1}^{n}\prod_{k=1}^{c_{ij}}\{I_j(t_{ijk})\,S_i(t_{ijk})\,p(t_{ijk};\,\mathbf{z}_i,\,\beta)\}^{\delta_{ijk}}\{1-I_j(t_{ijk})\,S_i(t_{ijk})\,p(t_{ijk};\,\mathbf{z}_i,\,\beta)\}^{1-\delta_{ijk}}. \quad (4)$$

We assume that the λ_{ij} are parameterized by $\theta=(\theta_1,\ldots,\theta_R)$ and that $p(t_{ijk};\,\mathbf{z}_i,\,\beta)=\exp(\beta\mathbf{z}_i)$, where β has length H. 0^0 is defined as 1. Assuming sufficient regularity such that the interchange of the various integrals and derivatives is justified, the $R+H$ score equations for level I can be written as

$$\frac{\partial\{\log L(C,N)\}}{\partial\theta_r}=\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^T\frac{1}{\lambda_{ij}(t)}\frac{\partial\lambda_{ij}(t)}{\partial\theta_r}\,dC_{ij}(t)-\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^T\frac{\partial\lambda_{ij}(t)}{\partial\theta_r}\,dt, \quad (5)$$

$$\frac{\partial\{\log L(C,N)\}}{\partial\beta_h}=\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^T\frac{I_j(t)\,S_i(t)z_{hi}}{1-\exp(\beta\mathbf{z}_i)}\,dN_{ij}(t)$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^T\frac{I_j(t)\,S_i(t)z_{hi}\exp(\beta\mathbf{z}_i)}{1-\exp(\beta\mathbf{z}_i)}\,dC_{ij}(t). \quad (6)$$

Since p lies in the interval [0, 1], in general we would want $\hat{\beta}\leqslant 0$. The score equations for β can be simplified. Let $\gamma_{ijk}=I_j(t_{ijk})\,S_i(t_{ijk})$ and

$$IC_{i.}=\sum_{j=1}^{n}\sum_{k=1}^{c_{ij}}\gamma_{ijk},$$

i.e. the total contacts made on person i by infectives while person i was susceptible. Making the above substitutions we obtain

$$\frac{\partial\{\log L(C,N)\}}{\partial\beta_h}=\sum_{i=1}^{n}\delta_i z_{hi}-\sum_{i=1}^{n}(IC_{i.}-\delta_i)\frac{z_{hi}\exp(\beta\mathbf{z}_i)}{1-\exp(\beta\mathbf{z}_i)}. \quad (7)$$

These equations are formally equivalent to a log-linear binomial regression where each person i with covariate $\mathbf{z}_i$ contributes $IC_{i.}$ trials with outcome δ_i. The score equations for β and θ can be solved separately. The information equations for this level and the score and information equations for all other levels are given in Rhodes *et al.* (1994a).

3.2. *Level II*

In level IIA the source of each infection is known, i.e. who infects whom, as well as how long each person is infectious. Level IIA is the last level with any direct contact information at all. On level IIB, it is known who is infectious and for how long, but no longer who infects whom. The time that a person remains infectious plus contact rates with other individuals gives a measure of the exposure opportunity that this person provides to other individuals, after taking into account when each was susceptible: level IIA,

$$\mathcal{H}_t^{\text{IIA}} = \sigma\{N_{ij}(s),\ I_j(s),\ S_i(s),\ \mathbf{Z}_i(s),\ \mathbf{G}_i(s),\ 0 \leqslant s \leqslant t\};$$

level IIB,

$$\mathcal{H}_t^{\text{IIB}} = \sigma\{N_{i.}(s),\ I_j(s),\ S_i(s),\ \mathbf{Z}_i(s),\ \mathbf{G}_i(s),\ 0 \leqslant s \leqslant t\}.$$

In most cases, information for pattern IIA will be difficult to obtain because of the necessity of observing who infects whom. When the C_{ij}-processes are not directly observed, we treat the N_{ij}-processes as thinned versions of the C_{ij}. Using expression (2) for the intensity of N_{ij}, the log-likelihood for level IIA can be written as

$$\log L(N_{ij}|S_i,\ I_j,\ \mathbf{Z}_i,\ \mathbf{G}_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} \delta_{ij} \log \lambda_{ij}(t_{ij}) + \sum_{i=1}^{n}\sum_{j=1}^{n} \delta_{ij}\boldsymbol{\beta}\mathbf{z}_i - \sum_{i=1}^{n} \exp(\boldsymbol{\beta}\mathbf{z}_i) \sum_{j=1}^{n} \int_0^T \lambda_{ij}(t)\, I_j(t)\, S_i(t)\, \mathrm{d}t. \tag{8}$$

Without knowledge of the contact process, we cannot estimate both the set of parameters λ_{ij} (or the $\boldsymbol{\theta}$) and the parameter β_0 corresponding to a constant term in $\mathbf{z}_i$. We must incorporate the value $\exp\beta_0$ into the λ_{ij}-functions and deal with a new set of parameters $\lambda_{ij}^* = \lambda_{ij} \exp\beta_0$. We shall also refer to the new set of parameters θ_1^* (note that $\theta_1^* \neq \theta_1 \exp\beta_0$ except in special cases). In this instance, the $\boldsymbol{\beta}$- and $\boldsymbol{\theta}^*$-equations cannot be solved separately. However, the score equations for $\boldsymbol{\beta}$ have the form of a Poisson regression if the terms involving the last portion of the second term, i.e.

$$\sum_{j=1}^{n} \lambda_{ij}^* \int_0^T I_j(t)\, S_i(t)\, \mathrm{d}t, \tag{9}$$

are known. Thus, estimation proceeds by alternating between solving the $\boldsymbol{\theta}^*$-equations and the $\boldsymbol{\beta}$-equations. Certain choices of the parameterization for the λ_{ij}^* lead to both sets of equations conforming to a Poisson regression model.

The intensities for the $N_{i.}$-processes are obtained by summing the intensities of the corresponding N_{ij}-processes (Bremaud, 1981). Level IIB has the same limitation in terms of not being able to estimate β_0 and λ_{ij} separately. Thus, the $N_{i.}$-processes have intensities

$$\alpha_{i.}(t) = \sum_{j=1}^{n} \alpha_{ij}(t) = \sum_{j=1}^{n} \lambda_{ij}^*(t)\, S_i(t)\, I_j(t) \exp(\boldsymbol{\beta}\mathbf{z}_i). \tag{10}$$

The log-likelihood for level IIB takes the form

$$\log L(N_{i.}\ i = 1, \ldots, n|I_j,\ S_i,\ \mathbf{Z}_i,\ \mathbf{G}_i) = \sum_{i=1}^{n} \delta_i \log\left\{\sum_{j=1}^{n} \lambda_{ij}^*(t_i)\, S_i(t_i)\, I_j(t_i) \exp(\boldsymbol{\beta}\mathbf{z}_i)\right\} - \sum_{i=1}^{n} \exp(\boldsymbol{\beta}\mathbf{z}_i) \sum_{j=1}^{n} \int_0^T \lambda_{ij}^*(t)\, I_j(t)\, S_i(t)\, \mathrm{d}t. \tag{11}$$

3.3. *Level III*

We know the times at which infections occur and which individuals were susceptible as well as the values of all covariate processes. We do not observe how long each person remains infectious. Thus, for level III,

$$\mathcal{H}_t^{\mathrm{III}} = \sigma\{N_{i.}(s),\ S_i(s),\ \mathbf{Z}_i(s),\ \mathbf{G}_i(s),\ 0 \leqslant s \leqslant t\}.$$

We proceed by writing a complete likelihood for the marked counting process $N_{..}(t) = \Sigma_{i=1}^{n} N_{i.}(t)$ and then decomposing it into components. The mark corresponds to the identity of the person infected when the combined process jumps. The contribution to the likelihood for the interval (t_{d-1}, t_d) where t_d is the time of the dth event in the process $N_{..}$ is broken into two parts:

(a) L(no event for $N_{..}$ in (t_{d-1}, t_d), event for $N_{..}$ at $t_d|\mathcal{H}_t^{\mathrm{III}}$, $t_{d-1} \leqslant t \leqslant t_d$);
(b) L(identity of person infected at t_d|event at t_d, set of individuals susceptible at time t_d, $\mathcal{H}_t^{\mathrm{III}}$, $0 \leqslant t < t_d$).

The first term is obtained by treating $N_{..}$ as the sum of thinned point processes and the second by considering the conditional probability of the identity of the infected individual given the set of individuals susceptible at time t_d. Level III has the same limitation in terms of not being able to estimate β_0 and λ_{ij} separately. The first term is equal to

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\{\lambda_{ij}^*(t_d)\, S_i(t_d)\, I_j(t_d) \exp(\beta \mathbf{z}_i)\} \exp\left\{-\int_{t_{d-1}}^{t_d} \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{ij}^*(t)\, S_i(t)\, I_j(t) \exp(\beta \mathbf{z}_i)\, dt\right\}, \tag{12}$$

whereas the second is given by

$$\frac{\exp(\beta \mathbf{z}_d) \sum_{j=1}^{n} \lambda_{ij}^*(t_d)\, I_j(t_d)}{\sum_{i=1}^{n}\left\{S_i(t_d) \exp(\beta \mathbf{z}_i) \sum_{j=1}^{n} \lambda_{ij}^*(t_d)\, I_j(t_d)\right\}}. \tag{13}$$

Thus, the conditional probabilities may depend on the contact parameters and on the I_j-processes. In some instances, depending on the form of the $\mathbf{G}_i$ covariates, strata can be formed in which the above conditional probability does not involve either the contact parameters or the I_j-processes. For example, if the $\lambda_{ij}(t)$ are all equal to a constant value λ, the conditional probability is free of both the above quantities. Also, consider the case where each individual belongs to one of K mixing groups. In that circumstance we can work with $N_{k..}$, $k = 1, \ldots, K$, the total infection processes in each of the K groups. Part (b) is then the conditional distribution of the mark given the actual set of individuals who were susceptible at time t_d in the group in which the infection occurred.

The Cox regression model has an advantage over analyses IIA and IIB in that no modification needs to be made for the situation where the study population constitutes only a portion of the entire population. For example, if we conduct a vaccine trial in a limited age group of the population and collect infection data only

for that age group, the Poisson-based methods could not be formulated correctly since we would not know the total exposure potential of the children in the trial.

Under a K-group mixing model, using a stratified analysis, the conditional probability will not depend on the unknown contact parameters (Rhodes *et al.*, 1994b). The Cox regression method will be useful only if this condition is met; otherwise the analysis still involves the contact parameters and the infectiousness processes. When this condition is met, the analysis is conducted by using only the second set of terms. Some information is lost by this strategy but usually only a small amount (Cox and Oakes, 1984).

The appropriate partial log-likelihood when the conditional probabilities of the marks do not depend on the contact intensities is

$$\log L_p = \sum_{i=1}^{n} \int_0^T \frac{\exp(\beta \mathbf{z}_i)}{\sum_{j=1}^{n} S_j(t) \exp(\beta \mathbf{z}_i)} \, \mathrm{d}N_{i\cdot}(t). \tag{14}$$

In a stratified analysis the value given above represents the contribution of one particular stratum.

3.4. *Level IV*

For level IV we know whether or not each individual has been infected in $(0, T]$ but not when the infection occurred:

$$\mathcal{H}_t^{\mathrm{IV}} = \sigma\{N_{i\cdot}(T), \mathbf{Z}_i(0), \mathbf{G}_i(0)\}.$$

The analysis has the form of a binary regression, although the link is the complementary log–log-link (i.e. $\log(-\log p)$). Censoring or late entry is not permitted; nor is it possible to incorporate time-dependent covariates. Thus, we restrict attention to the values of covariates at the start of the study.

Consider the probability that an individual i with covariates $\mathbf{z}$ would escape uninfected over the time period $(0, T]$ if we were given the full history of the infectiousness processes for all other individuals:

$$\Pr\{N_{i\cdot}(T) = 0 | I_j, \mathbf{Z}_i\} = 1 - p_i(T) = \exp\left\{ - \exp(\beta \mathbf{z}_i) \int_0^T \sum_{j=1}^{n} \lambda_{ij}(t) \, I_j(t) \, \mathrm{d}t \right\}, \tag{15}$$

or

$$\log[-\log\{1 - p_i(T)\}] = \beta \mathbf{z}_i + \log\left\{ \int_0^T \sum_{j=1}^{n} \lambda_{ij}(t) \, I_j(t) \, \mathrm{d}t \right\} = \beta \mathbf{z}_i + \gamma_i. \tag{16}$$

If the terms γ_i are unique to each individual, an estimation of the parameters of interest, β, is not possible, since each individual adds a new parameter to the analysis. However, if among the n individuals there is a limited number of γ-parameters, estimation is possible. Thus, although the I_j-processes are not observable, under certain conditions functions of these processes are estimable. However, these functions are not themselves of great interest. When there is a set of parameters

$\gamma = (\gamma_1, \ldots, \gamma_K)$, where $K \ll n$, we then fit the complementary log–log binomial regression model incorporating covariates for these parameters.

4. HOMOGENEOUS MIXING

We consider the case of homogeneous mixing, i.e. $\lambda_{ij}(t) = \lambda$ for $i \neq j$, with $p(t; \mathbf{z}_i, \boldsymbol{\beta}) = p_i = \exp(\beta_0 + \beta_1 z_i)$ for the case where z_i is a single dichotomous covariate. When the contact processes are not observable, the parameters λ and β_0 cannot both be estimated. The composite parameter $\lambda^* = \lambda \exp \beta_0$ is estimable and is interpretable as the average rate per unit of time at which one infective individual would tend to infect a susceptible individual with covariate equal to 0. The estimates for $\exp \beta_1$ for the different information levels and the corresponding estimated variances are given in Table 2. The estimator for level I has the form of a log-relative-risk. Analyses IIA and IIB are the same since there are no contact covariates. The estimator for β_1 for level II is similar to that for level I except that a measure of exposure opportunity is substituted for a measure of actual exposure. The Cox regression estimator (level III) does not have a closed form. The level IV estimator uses functions of the proportions infected in each group. If the probability of infection per contact is large, such as in measles or chicken-pox, analysis I might be a better choice than analysis II (Fig. 1). In this situation, knowledge of actual exposure, say a secondary attack rate study, provides a large improvement in the standard error over the use of expected exposure or exposure opportunity, such as a study using Poisson regression.

5. DISCUSSION

We have shown that the usual methods of analysis for estimating the effect of a covariate on susceptibility can be viewed as a hierarchy of parameters that depend on

TABLE 2

Estimates of β_1 and estimated variances for β_1 assuming homogeneous mixing†

Level	*Estimator*	*Variance estimator*
I	$\log\left(\frac{n_1 IC_0}{n_0 IC_1}\right)$	$\frac{1-\hat{p}_0}{n_0} + \frac{1-\hat{p}_1}{n_1}$
II	$\log\left(\frac{n_1 L_0}{n_0 L_1}\right)$	$\frac{1}{n_0} + \frac{1}{n_1}$
III	No closed form	No closed form
IV	$\log\left[\frac{\log\{-\log(1-\hat{p}_1)\}}{\log\{-\log(1-\hat{p}_0)\}}\right]$	$\sum_{i=0}^{1} \frac{\hat{p}_i}{m_i(1-\hat{p}_i)\{\log(1-\hat{p}_i)\}^2}$

†IC_i is the number of contacts made on individuals in group i by infectives while those individuals in group i were susceptible. n_i is the number of infections in each group during the study. L_i is the total time that susceptibles in group i were exposed to infectives. m_i is the initial number of susceptibles in group i, $\hat{p}_i = n_i/m_i$.

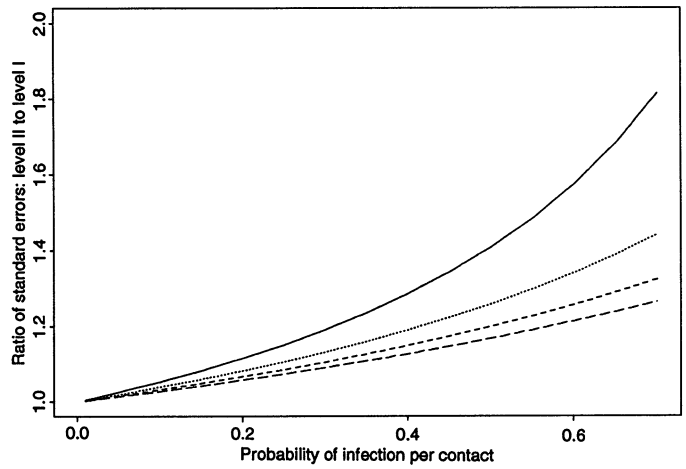

Fig. 1. Ratio of standard errors in the analysis at level II compared with level I by base-line transmission probability ($\exp\beta_0 = p_0$) and the covariate effect on the transmission probability, or transmission probability ratio (TPR $= \exp\beta_1$) in group 1 compared with group 0: ———, TPR = 1; ·····, TPR = 0.5; - - - - -, TPR = 0.25; – – – –, TPR = 0.1 (the ratios are based on the variances for β_1 at levels I and II given in Table 2; the number of infections is assumed to be the same in each group, and therefore cancels out)

how much information about the histories of the contact and infection processes is known or used in the analysis. We have used a marked counting process model for the combined contact and infection transmission processes that distinguishes the various roles played by variables affecting the amount of exposure to infection and variables affecting susceptibility to infection. A fundamental distinction is whether the measures require contact and exposure to infection information (Halloran and Struchiner, 1995).

When the contact process between infectives and susceptibles is observed, the analysis can be based on the relative transmission probability. When the contact process is not observable, the other measures make assumptions about equal exposure to infection in the comparison groups (Greenwood and Yule, 1915). Knowledge of the actual amount of exposure, measured by contacts with infectives, leads to a large gain in efficiency when the absolute probability of transmission per contact of an infective individual with a susceptible individual is high. Infectious diseases such as measles and chicken-pox have transmission probabilities greater than 0.85, whereas the transmission probability for the human immunodeficiency virus is generally less than 0.01, except perhaps during certain periods of infectiousness.

All the models with the exception of level IV can be extended to accommodate individuals who are lost to follow-up or who enter the population after the study starts. A more complicated situation is introduced by the process letting $Y_j(t) = 1$ if person j is *present* in the population at time t, and $Y_j(t) = 0$ otherwise. This differs from standard usage in survival analysis where $Y_j(t) = 1$ indicates that the person is *under observation* at time t (Andersen and Gill, 1982). A person who is not under observation but remains present in the population may influence the infection outcomes of other population members. This type of dependence is not seen in non-infectious disease studies (Ross, 1916).

Formulation of the contact process separately from the susceptibility allows us to study the bias in the estimates of relative susceptibility when exposure to infection is unequal in the comparison groups (Halloran *et al.*, 1994). It also allows a general formulation of the mixing between and within groups in terms of the between- and within-group contact rates. This enables an examination of the bias in the estimates of relative susceptibility when faulty assumptions are made about the mixing patterns (Rhodes *et al.*, 1994b). The derivations in this paper have assumed that the observed covariate value had an equal effect on all the people in one stratum. Unobserved heterogeneities, such as genetic variability, would have to be taken into account differently (Smith *et al.*, 1984; Svensson, 1991; Longini and Halloran, 1996; Halloran *et al.*, 1996).

We close with a guide to factors that may affect the choice of analysis. Knowledge of contacts between infectives and susceptibles, such as in secondary attack rate, case–contact or transmission studies (level I), provides the most efficient method of evaluating differential susceptibility. Analyses IIA and IIB are useful if we are interested in studying the extent to which different segments of the population spread infection as well as the effects of susceptibility variables. If the appropriate data are available, level IIA provides a more efficient and more easily implemented analysis of such structure. If the mixing structure is not of inherent interest, the Cox regression method (level III), when appropriate, is an alternative which is easier to implement. The Cox model is useful in situations where we are not confident about information or assumptions concerning how long individuals remain infectious. Finally, if we know only the number of events that occur over the course of a study, then level IV, the complementary log–log-model, provides an efficient alternative for a closed population with moderate overall levels of infection.

ACKNOWLEDGEMENTS

This work was partially supported by National Institute of Allergy and Infectious Diseases awards R29-AI31057 (to MEH) and R01-AI32042 (to MEH and IML).

REFERENCES

Andersen, P. K. and Gill, R. D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, **10**, 1000–1120.

Arjas, E. (1989) Survival models and martingale dynamics. *Scand. J. Statist.*, **16**, 177–225.

Becker, N. G. (1982) Estimation in models for the spread of infectious diseases. In *Proc. 11th Int. Biometrics Conf.*, pp. 145–151. Versailles: Institut National de la Recherche Agronomique.

——— (1985) A generalized linear modeling approach to the analysis of a single epidemic. In *Proc. Pacific Statistical Congr.* (eds I. Francis, B. Manley and F. Lam), pp. 464–467. Amsterdam: North-Holland.

——— (1989) *Analysis of Infectious Disease Data.* London: Chapman and Hall.

Bremaud, P. (1981) *Point Processes and Queues.* New York: Springer.

Cox, D. R. and Oakes, D. (1984) *Analysis of Survival Data.* London: Chapman and Hall.

Fleming, T. and Harrington, D. (1991) *Counting Processes and Survival Analysis.* New York: Wiley.

Greenwood, M. and Yule, U. G. (1915) The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. *Proc. R. Soc. Med.*, part 2, **8**, 113–194.

Halloran, M. E., Longini, I. M. and Struchiner, C. J. (1996) Estimability and interpretation of vaccine efficacy using frailty mixing models. *Am. J. Epidem.*, to be published.

Halloran, M. E., Longini, I. M., Struchiner, C. J., Haber, M. J. and Brunet, R. C. (1994) Exposure efficacy and change in contact rates in evaluating prophylactic HIV vaccines in the field. *Statist. Med.*, **13**, 357–377.

Halloran, M. E. and Struchiner, C. J. (1995) Causal inference for infectious diseases. *Epidemiology*, **6**, 142–151.

Longini, I. M. and Halloran, M. E. (1996) A frailty mixture model for estimating vaccine efficacy. *Appl. Statist.*, **45**, 165–173.

Rhodes, P. H., Halloran, M. E. and Longini, I. M. (1994a) Counting process models for differentiating exposure to infection and susceptibility. *Technical Report 94-1*. Division of Biostatistics, Emory University School of Public Health, Atlanta.

———(1994b) Analysis of susceptibility to infection with K-group mixing models using counting processes. *Technical Report 94-9*. Division of Biostatistics, Emory University School of Public Health, Atlanta.

Ross, R. (1916) An application of the theory of probabilities to the study of *a priori* pathometry, part 1. *Proc. R. Soc.* A, **92**, 204–230.

Smith, P. G., Rodrigues, L. C. and Fine, P. E. M. (1984) Assessment of the protective efficacy of vaccines against common diseases using case–control and cohort studies. *Int. J. Epidem.*, **13**, 87–93.

Svensson, Å. (1991) Analyzing effects of vaccines. *Math. Biosci.*, **107**, 407–412.