

Statistical Interpretation of Data from Partner Studies of Heterosexual HIV Transmission

Stephen Shiboski*
University of California, San Francisco

*Department of Epidemiology and Biostatistics; San Francisco General Hospital; Bldg 90, Ward 95,
Room 512; 995 Potrero Avenue; San Francisco, CA 94110; (415)476-5325

1 Introduction

Since the beginning of the AIDS epidemic, epidemiologists have sought to understand biological and behavioral factors influencing transmission of the Human Immunodeficiency Virus (HIV) from infected to susceptible individuals. Knowledge of these factors is critical in understanding the mechanisms of transmission and in developing methods to control and forecast the spread of HIV. In addition to identifying risk factors, epidemiological investigations seek to quantify the risk of virus transmission associated with these factors. Of particular interest are transmission probabilities (probability of virus transmission given that exposure has occurred) and the per-contact transmission risk or *infectivity*.

A major difficulty in epidemiological studies of sexual transmission is linking observed behavioral and biological data to transmission risk. This is mainly due to the fact that data come almost exclusively from cross-sectional samples of previously infected and uninfected individuals - all information on sexual behavior must be retrospectively ascertained. In assessing transmission risk, infected individuals are compared to the uninfected with respect to potential risk factors. For factors such as frequency and type of contact, it is often impossible to know whether or not reported behaviors truly represent exposure (i.e. unprotected contact with an infected person). Further, when multiple partners are reported, it is difficult or impossible to verify whether they were infected or not. Consequently, estimated quantities such as transmission probabilities, infectivity and measures of association are either indeterminate, or are subject to substantial bias.

Partner studies of sexual transmission are designed to avoid some of these difficulties by limiting enrollment to monogamous partnerships with a clearly defined primary infected person, often termed the *index case*. Exposure is therefore clearly defined as contact with a known infected person, and estimates of the frequency of exposure are readily obtained via interview. Despite these advantages, many problems with data interpretation remain. In particular, if the time of index case infection is unknown, the length of exposure cannot be accurately determined.

The major objective of this chapter is to provide a framework for assessing validity of conclusions drawn from epidemiological studies of HIV transmission. Section 2 contains a brief description of the basic design features of partner studies of HIV transmission, Section 3 reviews recently developed techniques for the analysis of partner study data [1, 2], and Section 4 provides quantitative descriptions of several sources of bias which arise in the analysis of such data. The results are applied to data from the California Partners' Study and the CDC Transfusion study of heterosexual transmission in Section 5. The major analytical results indicate that statistical interpretation of partner study data is hampered by a number of factors, including biased selection of partnerships, measurement error in exposure data and unobserved variables related to infectiousness and/or susceptibility. These factors introduce bias into estimates of transmission probabilities and regression coefficients, and can lead to incorrect inferences regarding risks of heterosexual HIV transmission. Analyses of example datasets indicate that although the assumption of constant infectivity appears to be false, poten-

tial bias due to the factors listed above precludes obtaining any detailed information about the shape or scale of the infectivity from partner study data.

2 Data from Partner Studies

This section introduces design issues and data collection schemes for partner studies. A fuller treatment of these issues is provided in [3] and [4].

A partner study consists of a number of partnerships, each made up of a primary infected partner or *index case*, and a secondary partner (also referred to as the *partner*) who may or may not be infected at the time the couple is recruited for study. Eligibility requirements include verification that the primary partner was infected first, and that the only source of exposure to HIV for the secondary partner is through sexual contact with the primary partner. On recruitment, data on the partnership is collected, including frequency and type of sexual contacts, length of exposure of the partner to the index case, time of infection of the index case (if known) and current serostatus of the partner. A *contact* will be defined here to be a single act of unprotected sexual intercourse; *exposure* refers to the contact history of a particular partnership, and includes the frequency of contacts, and the length of time over which the contacts occurred following infection of the index case. The primary random variables summarizing exposure are defined symbolically below:

$$Y = \begin{cases} 1 & \text{if both partners are infected at recruitment} \\ 0 & \text{if only one partner is infected at recruitment,} \end{cases}$$

V = time between index case infection and partner infection (unobserved),

T = length of exposure interval,

M = rate of contacts occurring in exposure interval,

K = number of exposed contacts ($K = M \times T$).

These variables are shown graphically in Figure 1. Contacts are routinely measured as a rate M per unit time of exposure. Because the data are collected retrospectively, this rate is usually assumed to be constant. As indicated in the definitions above, contact counts can be estimated from the rates and corresponding exposure durations by multiplication.

A major goal in statistical analysis of such data is to understand how transmission probabilities and infectivity depend on exposure. If transmission risk varies with duration and degree of exposure, then these factors need to be accounted for in assessing the effects of other risk factors. Of particular interest are characteristics of the length of time V between infection of the index case and transmission to the partner. In survival analysis terminology this can be regarded as a "failure" or survival time, and features of its distribution such as the probability of transmission given a fixed duration of exposure (the transmission probability) and the "instantaneous" probability of transmission in a very small time interval (the infectivity) of exposure provide key measures linking exposure to transmission risk.

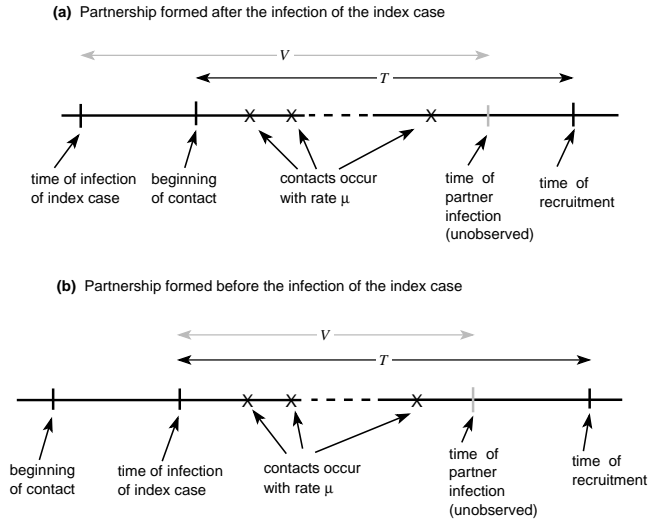


Figure 1. Illustration of exposure data collected in retrospective partner studies. The upper diagram (a) displays data from a short term partnership, in which contact begins after the infection of the index case. Data from a long term partnership, in which contact begins before the infection of the index case, is displayed below (b).

Figure 1 makes it clear that due to the retrospective nature of data collection, V is never actually observed in partner study data. However, observations of the infection status Y and exposure duration T at recruitment provide information about V which can be used to investigate properties of V indirectly. Figure 1 illustrates two basic types of partnerships with regard to exposure history: *long-term* partnerships, in which the partnership is known to have started prior to infection of the index case, and *short-term* partnerships, in which the partnership is known to have started after infection of the index case. The figure shows clearly that if the time of infection of the index case is completely unknown, then accurate quantification of exposure duration and intensity are impossible. For short-term partnerships, all contacts can be considered potentially exposed, while for long-term partnerships, only those contacts occurring after index case infection carry transmission risk. Frequently it is only possible to determine that the index case was infected within a broad interval, possibly extending back to the beginning of the epidemic. In such cases, precise knowledge of the duration and degree of exposure is impossible and the beginning of exposed contact must be estimated. For long-term partnerships this is somewhat easier, since in many cases contact was known to have started before the earliest possible date of index case infection. For this reason, and to simplify the presentation of ideas, the remainder of the chapter will focus primarily on long-term partnerships.

3 Statistical Methods

This section reviews methods for statistical analysis of partner study data. The development is primarily based on references [1] and [2].

3.1 Transmission Models

Unlike diseases with completely unknown etiologies, the agents of sexually transmitted diseases such as AIDS are often known and the mechanisms of transmission somewhat understood. Thus, even a greatly oversimplified mathematical description of transmission can capture essential elements of the transmission process and provide useful insights into factors influencing disease spread. For this reason, epidemiologists often rely on simple models of transmission to assist in understanding the relative importance of factors influencing exposure, infectiousness and susceptibility.

3.1.1 The Constant Infectivity Model

The simplest model for sexual transmission of HIV is based on the probability that a susceptible individual is infected following a specified number of contacts with an infected person. Recalling the notation from the previous section, suppose the i_{th} partner of a sample of n partnerships is exposed to the index case for t_i time units following the latter's infection, and that contacts occurred at constant rate μ_i during this period. Here, t_i and μ_i refer to the observed values of the random variables T_i and M_i , respectively. Then, if the per-contact infection risk or infectivity is λ and contacts occur independently, the probability that the partner escapes infection is

$$\text{pr}(Y_i = 0 \mid k_i) = (1 - \lambda)^{k_i}, \quad (1)$$

where k_i gives the estimated number of exposed contacts (defined above).

Jewell and Shiboski [1] develop methods for estimating λ based on partner study data by observing that a transformation of the above expression yields a standard generalized linear regression model [5] for the binomially distributed outcome Y . Denoting the above probability by $S(k_i)$,

$$\log[-\log S(k_i)] = \log[-\log(1 - \lambda)] + \log k_i .$$

In the terminology of generalized linear models, the complementary log-log transformation is the correct *link function* to produce a linear additive relationship between transmission probability and exposure. The intercept term of this linear model depends on the infectivity, and the "independent" variable is $\log k_i$ with unit slope.

A key feature of the constant model is that it does not depend on duration of exposure - transmission depends only on the number of contacts and the placement of these in the exposure interval is ignored. Clearly if transmission risk varies with length of exposure (e.g. transmission is more likely as the index case progresses to a more advanced stage of disease), the model omits an important aspect of the transmission process.

3.1.2 Time Dependent Infectivity

To address the major limitation in the constant infectivity model raised above, Shiboski and Jewell [2] derived the following expression for the probability that a partner remains infection free following an exposure interval of length t_i , with contact rate and infectivity at time x given, respectively by $\mu_i(x)$ and $\lambda(x)$:

$$\begin{aligned} S(t_i | \mu_i(x); 0 \leq x \leq t_i) &= \text{pr}(Y_i = 1 | t_i, \mu_i(x); 0 \leq x \leq t_i) \\ &= \exp \left\{ - \int_0^{t_i} \lambda(x) \mu_i(x) dx \right\} . \end{aligned} \quad (2)$$

In the language of survival analysis [6] this equation specifies a *survival function*. The transmission probability is obtained from the survival function by subtraction from one. The associated *hazard function* gives the probability of transmission occurring in a very short interval conditional on being uninfected at the start of the interval, and is defined by

$$\begin{aligned} h(v | \mu_i(x), 0 \leq x \leq t_i) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \text{pr}\{v \leq V \leq v + \Delta | V \geq v; \mu_i(x), 0 \leq x \leq t_i\} \\ &= \lambda(v) \mu_i(v) . \end{aligned} \quad (3)$$

Note that the infectivity and contact rate functions appear together in this expression, indicating that the infectivity function cannot be separately estimated unless the contact rate function is known. In the rest of the chapter, this hazard function will be referred to as the *infection hazard* function and the contact rate function will be assumed to be well approximated by a constant for each partnership.

Model (2) rests on the assumption that contacts occur according to a non-homogeneous Poisson process with rate $\mu(\cdot)$. In practice, due to the retrospective nature of partner study data, the rate of contacts is assumed to be constant, leading to the simplified model

$$\begin{aligned} S(t_i | \mu_i) &= \exp \left\{ -\mu_i \int_0^{t_i} \lambda(x) dx \right\} \\ &= S_0(t_i) \exp(\log \mu_i) . \end{aligned} \quad (4)$$

This specifies a proportional hazards relationship [6] for the dependence of the distribution of the unobserved time of partner infection V on the rate of contacts μ . The term $S_0(t_i)$ in (4) represents the "baseline" probability of the i th partner remaining infection free following an exposure of length t_i , with unit contact rate. The infection hazard function for this model can be written

$$h(t_i | \mu_i) = \lambda(t_i) e^{\log \mu_i} ,$$

which is clearly of the proportional hazards form.

Just as in the constant infectivity case, equation (4) can be transformed to a linear additive form:

$$\log [-\log S(t_i | \mu_i)] = \log \int_0^{t_i} \lambda(x) dx + \log \mu_i . \quad (5)$$

The intercept term in this equation is a monotone increasing function of length of exposure. If no restrictions are made on the form of the infectivity function $\lambda(\cdot)$ the equation defines a *semiparametric* binary regression model [7] and is also (due to the monotonicity constraint on the intercept term) an example of a binary additive isotonic regression model [8]. This additive linear form, coupled with the proportional hazards link between exposure and outcome provides a convenient framework for estimation and inference. In particular, additional covariates can easily be included on the right hand side of (5).

In the case that the infectivity function is assumed to be constant ($\lambda(x) \equiv \lambda, x \geq 0$), this model becomes

$$\log[-\log S(t_i | \mu_i)] = \log \lambda + \log \mu_i t_i, \quad (6)$$

in which $S(t_i | \mu_i)$ is the time based analog of equation (1). Because these versions of the constant infectivity model are almost indistinguishable for small values of λ (i.e. $-\log(1 - \lambda) \approx \lambda$ for small λ), the remainder of the chapter will focus on (6).

3.1.3 Heterogeneous Infectivity

Implicit in both the constant and time dependent infectivity models presented above is the assumption that infectivity is a deterministic quantity shared by all partnerships. Thus, no accommodation is made for partnership-specific variation in infectiousness and/or susceptibility. Measured covariates can be expected to account for some of this variation. However, due to the retrospective nature of partner study data it is unlikely that all relevant variables can be observed. Thus, partnership heterogeneity in transmission risk can be viewed as arising from the influence of unmeasured covariates.

One way to introduce such heterogeneity into these models is to assume that the infectivity is a random quantity with known distribution (i.e. that each partnership is allowed to have a unique infectivity). For example, assume that the individual-specific infectivity for the i th partnership in a sample of n , with length of exposure t_i is given by $\lambda_i(t_i) = W_i \lambda(t_i)$, where W is a non-negative random variable with unit expectation and distribution function $K_W(w)$. This is an example of a *mixture model* [9], in which W represents the individual heterogeneity in infectivity (i.e. the mixing variable), and the function $\lambda(\cdot)$ represents the average infectivity over the population. Because the random variable W is unobserved, estimation and inference are based on the following marginal survival function:

$$\bar{S}(t | \mu) = \int_0^\infty \exp \left\{ -w \mu \int_0^t \lambda(x) dx \right\} dK_W(w | t, \mu). \quad (7)$$

Note that the mixing variable W in this model could also be interpreted as multiplicative measurement error associated with the covariate μ , or as representing the multiplicative effect on the infectivity of omitted covariates. This illustrates the close link that exists between models for assessing the impact of uncontrolled heterogeneity, measurement error and unobserved covariates.

In practice, the distribution of W is usually assumed not to depend on covariates such as t and μ . Although this model makes a strong assumption about how heterogeneity impacts the infectivity, it provides a simple means of evaluating the potential impact of heterogeneity on estimated transmission probabilities, infectivities and regression coefficients. These issues will be investigated further in Section 4.1.

3.2 Estimation

Maximum likelihood-based methods can be used to obtain parameter estimates for all transmission models considered above. This subsection covers estimation techniques for data from partner studies consisting exclusively of long-term partnerships. Shiboski and Jewell [2] present techniques which accommodate short-term partnerships as well.

Recall from Section 2 that the observed data for a long-term partnership consist of the triplets (y, t, μ) - the actual time V to partner infection is unobserved. In the language of survival analysis, all observations of V are *censored*, either on the left for partners observed to be infected at recruitment ($y = 1$), or on the right for partners who are uninfected at this time ($y = 0$). Figure 1(b) illustrates the definitions. This is a form of *interval censoring*, also known as *current status data* [10], since the current infection status of partners is the only data on transmission available at the time couples are enrolled into a partner study.

The likelihood for n partnerships, each yielding observations of the form (y_i, t_i, μ_i) is given by

$$\prod_{i=1}^n [1 - S(t_i | \mu_i)]^{y_i} S(t_i | \mu_i)^{1-y_i}, \quad (8)$$

where the first term reflects the contribution of a partnership in which the partner has been infected, and the second, the contribution from a partnership with an uninfected partner. This likelihood is based only on the distribution of V *conditional* on the observed values of the duration of exposure t and the contact rate μ , the joint distribution of which is assumed to provide no additional information about the form of $S(t_i | \mu_i)$.

The log-likelihood associated with (8) can be minimized by substituting in the appropriate expression for $S(t_i | \mu_i)$, and applying standard iterative maximum likelihood techniques. For the constant model (6), standard programs for generalized linear models can be used, employing binomial errors for the response Y , and the complementary log-log link function [5]. In addition, Kaplan [11] presents an approximate, closed form approach to estimating λ in (6). Estimation for the general time dependent model (5) is more difficult due to the facts that the intercept term is constrained to be monotone increasing in t , and that no form is assumed for the infectivity function $\lambda(\cdot)$. More details on estimation are provided in [2] and [4].

3.3 Inference

Statistical inferences drawn from partners study data focus on three major issues: (i) evaluation of the hypothesis of constant infectivity, (ii) uncertainty in estimates of

transmission probabilities and infectivity and (iii) investigation of the dependence of transmission probabilities on risk factors, including the assessment of confounding and interaction for multiple factors.

3.3.1 Evaluating the Constant Infectivity Assumption

Recent biological evidence on shedding of HIV in infected persons indicates that infectiousness may vary with disease stage, with periods of increased infectiousness coming early and again late in the course of infection [12]. In addition, susceptibility may vary during the course of a partnership. These factors may lead to a systematic change in infectivity on the time scale measured from time of infection of the index case, and invalidate the assumption of constant infectivity. This assumption is central to many mathematical models which have been developed to predict the spread of HIV, and is implicit in many analyses of epidemiological data on heterosexual HIV transmission. Thus, evaluation of the validity of this assumption is an important step in data analysis.

As shown in [1], a preliminary assessment of the fit of the constant infectivity model can be made by fitting the following version of (6):

$$\log[-\log S(t | \mu)] = \theta + \gamma \log \mu t . \quad (9)$$

An estimated coefficient $\hat{\gamma} < 1$ may indicate a systematic departure from constant infectivity which may be due to one or more of the following factors: time dependence in the infectivity function or contact rates, heterogeneity of infectivity across partnerships, measurement error in exposure information and biased selection of partnerships in the sample. (Although attenuation of regression coefficients from their true values is the most common effect of these factors, in Section 4 it will be shown that under certain conditions they may also lead to inflation of coefficient estimates.) Setting $\theta = \log \lambda^\gamma$ in (9), the infection hazard function associated with the survival function defined in that equation can be written:

$$h(t | \mu) = -\frac{d}{dt} \log S(t | \mu) = \gamma \lambda e^{\log \mu} (\lambda e^{\log \mu} t)^{\gamma-1} ,$$

which is a hazard function for the Weibull distribution provided that γ is positive [6]. For $\mu = 1$, this equation gives a representation for the transmission risk associated with a single contact (i.e. the infectivity). For $\gamma = 1$, it reduces to (6). However, the link between the covariate $\log \mu$ and time in the above infection hazard function is not of the proportional hazards form specified in (5) and (6). Thus, an observed improvement in fit of (9) over (6) is an indication that the proportional hazards assumption may not hold. Results in Section 4 show that this lack of fit may be symptomatic of a number of factors (mentioned above) other than time dependence, all of which lead to observed departures from constant infectivity.

A more general approach to evaluating time dependence involves fitting model (5), which makes no parametric assumptions about the form of the infectivity, and comparing the fit to (6). Because the estimates of the infectivity function in (5) are nonparametric, measures of uncertainty and associated inference procedures are not well developed

(Groeneboom and Wellner [13] provide details about what is currently known). Simulation tests or bootstrap procedures must therefore be used to evaluate the improvement in fit over the constant model. Lack of fit of (5) can be assessed using the same means proposed for the constant model - by fitting a model analogous to (9)

$$\log[-\log S(t|\mu)] = \log \int_0^t \lambda(x) dx + \gamma \log \mu, \quad (10)$$

and looking for attenuation in $\hat{\gamma}$. As before, an estimated coefficient $\hat{\gamma} < 1$ calls into question the postulated form of the model (5), and may indicate uncontrolled heterogeneity of infectivity across partnerships, measurement error in exposure information and biased selection of partnerships in the sample. In addition to their use in investigating departures from constant infectivity, the nonparametric estimates are useful in suggesting possible parametric forms for the infectivity which, if suitable, may be supported by conventional inference for parametric models.

Section 4 identifies several potential sources of bias in partner study data, and investigates their effects on estimates of transmission probabilities, regression coefficients and infectivity. In addition, some simple procedures for detecting the presence of bias are proposed.

3.3.2 Investigating Covariate Effects

The influence of additional covariates on transmission risk can easily be investigated in the models introduced above. For example, model (5) (or (10)) can be extended to incorporate a vector of covariates \mathbf{z} , with associated parameter vector β as follows:

$$\log[-\log S(t|\mu, \mathbf{z})] = \log \int_0^t \lambda(x) dx + \log \mu + \beta^T \mathbf{z}. \quad (11)$$

Standard methods of inference from generalized linear models [5] can then be applied to address questions on the relationship of transmission risk to additional covariates, and the issues of confounding and interaction.

4 Sources of Bias in Partner Study Data

This section investigates bias which can arise in analyses of partner study data. Although the sources of bias discussed are not unique to partner studies, they can be particularly severe for such studies due to the retrospective nature of data collection, uncertainty about key exposure events, and to the difficulties associated with recruitment discussed in Section 2. The discussion here focuses on assessing bias in the framework of the time-dependent infectivity model (5), and quantitative assessment of the effect of various sources of bias on estimated transmission probabilities and regression coefficients. The major goal is to provide guidance on when quantitative conclusions drawn from partner study data can be considered reliable. Similar issues are discussed in the context of the constant infectivity model in [1] and [4]. In addition

to influencing parameter estimates, the sources of bias discussed here can also affect estimates of the variances associated with the estimates. This topic is not covered in this chapter.

4.1 Neglected Heterogeneity of Infectivity

As discussed in Section 3.1.3, the assumption of a deterministic infectivity function common to all partnerships made in models (5) and (6) may be overly restrictive in view of recent biological and epidemiological evidence that infectiousness may vary with disease stage of the index case. Given the limitations inherent in available partner study data, the goal of accurately modeling and estimating a random infectivity is unrealistic. This raises the question of whether simple (deterministic) transmission models can still be used to investigate the effects of measured risk factors in the presence of uncontrolled heterogeneity, and about the nature of the bias in infectivities estimated from such models. Jewell and Shiboski [1] evaluated the impact of neglected heterogeneity of infectivity on estimated regression coefficients in the context of the constant infectivity model (1). Their results indicate that fitting such a model when the underlying infectivity is random leads to attenuated estimates of regression coefficients. The magnitude of this attenuation was calculated for the case where the (constant) infectivity varied between partnerships according to a beta distribution. They also showed that the estimated (constant) infectivity was an overestimate of the median of the distribution of the true random infectivity. Shiboski and Jewell [2] speculated that similar results would hold for the time-dependent infectivity model (5). Here, the random infectivity model introduced in Section 3.1.3 is used to investigate these issues further.

A number of authors have investigated the impact of individual heterogeneity on estimation and inference in survival analysis [9, 14]. Their results show that the presence of heterogeneity leads to distortions in estimated hazard functions, survival probabilities and regression coefficients. If heterogeneity is assumed to be introduced into the infectivity function via the multiplicative form introduced in Section 3.1.3, then these results can be used to quantify these forms of bias in the context of the transmission models considered here. Recall from Section 3.1.3 that $\bar{S}(t | \mu)$ denotes the marginal survival function corresponding to observed data from a partnership with exposure of duration t and with contact rate μ , where the true survival function depends on the unobserved variable random variable W . Define

$$\bar{h}(t | \mu) = -\frac{d}{dt} \log \bar{S}(t | \mu) ,$$

the associated marginal infection hazard function. It is shown in Appendix A.1 that

$$\bar{h}(t | \mu) = \mu \lambda(t) C , \tag{12}$$

where C is a constant depending on t , and that

$$\frac{d}{dt} \log \bar{h}(t | \mu) < \frac{d}{dt} \log \mu \lambda(t) . \tag{13}$$

These results indicate that, given the multiplicative model for heterogeneity and that contact rates are constant, that the infectivity estimated if heterogeneity is ignored differs in scale and rate of decline from the average infectivity $\lambda(t)$. Thus, both scale and shape of the estimated infectivity are distorted due to neglected heterogeneity. For example, assume that the underlying infectivity is constant and that W follows a gamma distribution with unit expectation and squared coefficient of variation δ . In this case,

$$\bar{h}(t | \mu) = \lambda\mu / [1 + \delta\mu\lambda t], \delta \geq 0.$$

This equation shows that for fixed contact rate μ , the estimated infection hazard will appear to vary in time and will be smaller than the average infection hazard, with the discrepancy increasing with both length of exposure and degree of variability in the distribution of W . Figure 2 illustrates this phenomenon for different values of δ assuming that $\mu = 1$.

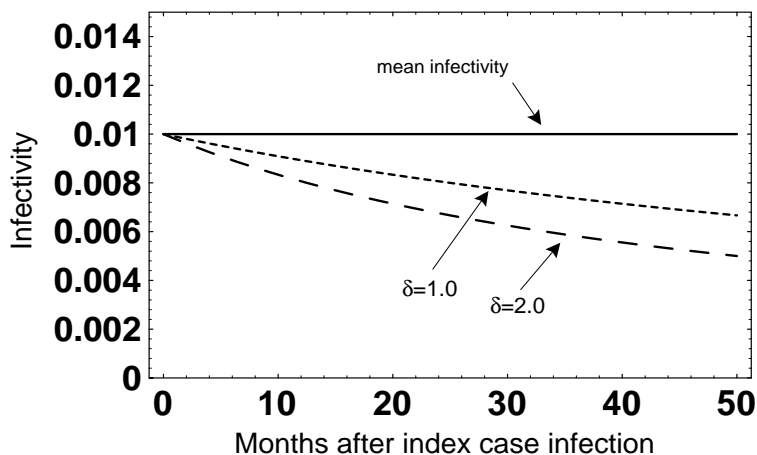


Figure 2. Mean infectivity (solid line) and marginal infectivity function (dashed lines) from gamma mixing model for two choices of the squared coefficient of variation δ .

In practice, a preliminary assessment of the presence of uncontrolled heterogeneity of infectivity can be obtained from fitting the time dependent models (9) or (10). As discussed above, an estimated coefficient $\hat{\gamma} < 1$ may indicate the presence of such heterogeneity. Similarly, if additional covariates are present in the model (e.g. (11)), the associated coefficients will be attenuated as well. If heterogeneity is introduced according to the multiplicative model (7) and the distribution of the mixing random variable W is assumed to be independent of μ and t , the attenuation in regression coefficients such as $\hat{\gamma}$ can be explicitly calculated, and is shown in Appendix A.1 (equation (22)) to have the form

$$\left\{ 1 - \left[\int_0^t \mu\lambda(s)ds \right] \frac{\text{Var}[W | V \geq t]}{\text{E}[W | V \geq t]} \right\}.$$

Taking models (9) or (10) as an example, the estimated regression coefficient $\hat{\gamma}$ in the presence of heterogeneity is strictly less than the true value of $\gamma = 1$ in (5) (or (6)). Figure 3 illustrates this phenomenon for the special case of the constant infectivity model (9), assuming that W follows a gamma distribution. Note that the multiplicative factor which describes the attenuation depends on the underlying infectivity and the duration of exposure, as well as the conditional distribution of the random variable W among the individuals who have escaped infection up to time t . Thus, when the data are affected by uncontrolled heterogeneity, the proportional hazards assumption implicit in transmission models (9) and (10) *will not hold*. Further, if the distribution of W depends on covariates such as μ and/or the duration of exposure t as specified in (7), then the relationship between the estimated coefficient $\hat{\gamma}$ from models (9) and (10) may differ in both sign and magnitude from the true value of unity (see Appendix A.1). In this case reliable inferences about the influence of covariates on transmission risk are impossible.

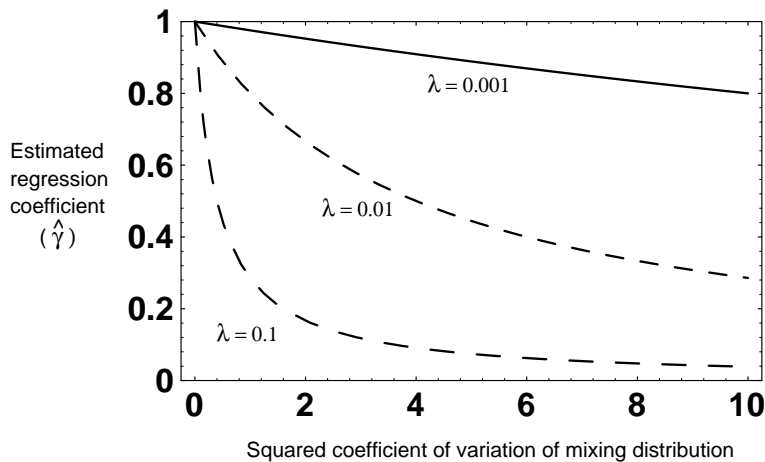


Figure 3. Attenuation in estimated value of coefficient γ obtained from fitting model (9) when the gamma mixing model applies, plotted as a function of the squared coefficient of variation δ from the mixing distribution and three values of the underlying infectivity λ .

A number of factors other than heterogeneity may result in an estimated coefficient $\hat{\gamma} < 1$ for models (9) and (10). Thus, a more direct approach to investigating heterogeneity is to fit parametric versions of the mixture model introduced in Section 3.1.3. However, partner studies often have very small sample sizes with observations limited to one per partnership. Such data often lead to unstable or indeterminate estimates of mixing distribution parameters. An alternative to fitting specific mixture models is to construct tests to detect whether the observed data are *overdispersed* relative to the assumed (binomial) distribution for the data. Cox [15] proposed a method for testing for

overdispersion which involves forming an approximation to the distribution generating the data, assuming that the degree of overdispersion is small. Lancaster [16] applied Cox's ideas to mixture models of the form introduced in Section 3.1.3. In the present context, this procedure involves assuming that the data arise from a mixture model of the form presented in Section 3.1.3, with conditional survival function

$$S(t | \mu, W) = \exp \left[-W \int_0^t \mu \lambda(x) dx \right],$$

where W is the mixing random variable with unit expectation and variance σ^2 . As shown in Appendix A.2, expanding this function in a second order Taylor series about $\sigma^2 = 0$, and taking the expectation with respect to W leads to the following approximation to the marginal survival function:

$$\bar{S}(t | \mu) \approx \exp[-M(t)] \left\{ 1 + \frac{\sigma^2}{2} [M(t)]^2 \right\},$$

where $M(t) = \int_0^t \mu \lambda(x) dx$. When $\sigma^2 = 0$, corresponding to the absence of heterogeneity, this equation reduces to model (4). This suggests that a test of the hypothesis that $\sigma^2 = 0$ will provide a means of detecting overdispersion in the data without assuming a form for the distribution of the mixing variable W . If parametric assumptions are made about the infectivity (e.g. constant) standard likelihood methods can be used to construct a test statistic with known distributional properties. Details of a score test statistic for this hypothesis are provided in Appendix A.2.

4.2 Biased Recruitment of Partnerships

For reasons given in Section 2, few partnerships meet the strict eligibility requirements for partner studies. In most cases, couples are enrolled on a volunteer basis, or recruitment is targeted at known high-risk populations. Consequently, the representativeness of a sample of partnerships is often in doubt. *Selection bias* arises in a partner study when the probability of enrolling a partnership in which transmission has occurred differs from the analogous probability for partnerships with uninfected partners. For example, in studies which recruit volunteers, couples with infected partners who are aware of their serostatus may be more likely to enroll than couples in which partners are ignorant of their serostatus. Studies where recruitment is based on identifying index cases by AIDS diagnoses provide another example. In these studies, only index cases who have developed AIDS in the enrollment period will be recruited. These cases will usually represent only a subset of potential AIDS cases, and may be distinguished by shorter incubation periods, particular clinical manifestations, or other factors which depend on the chronological placement of the enrollment period. If these characteristics (e.g short incubation length) are in turn related to HIV transmission probability, a biased sample will clearly result from this form of partnership selection.

The impact of selection bias on estimated coefficients from binary regression models has been examined extensively for the logistic model. It is well known [5] that unequal selection probabilities under retrospective sampling do not influence regression

coefficient estimates for covariates in this model, provided that selection probabilities themselves do not depend on the covariates. This is not true for binary regression models in which the response probabilities are linked to the covariates with functions other than the logit, such as the complementary log-log link function. All the transmission models (e.g (5), (6), (9) and (10)) considered here are based on this link function, which implies that estimates of both the intercept term and regression coefficients will be biased in the presence of differential selection probabilities. It follows that estimates of transmission probabilities and the infectivity will also be biased.

For the time dependent infectivity model (5), it is shown in Appendix B that under selection bias, the estimated probability of a partner being observed to be uninfected following t units of exposure (denoted $S^*(t | \mu)$) has the following relationship to the true survival function $S(t | \mu)$:

$$S^*(t | \mu) = \frac{\alpha(t, \mu)S(t | \mu)}{1 - S(t | \mu) + \alpha(t, \mu)S(t | \mu)}. \quad (14)$$

The term $\alpha(t, \mu)$ represents the ratio of the probability of selecting a partnership with an uninfected partner to the probability of selecting a partnership with an infected partner. In the general case, these probabilities can depend on both duration of exposure t and on covariates such as the contact rate μ . Examination of (14) shows that when selection bias is operating (i.e. when $\alpha(t, \mu) \neq 1$), the estimated proportion of partners escaping infection may be underestimated or overestimated, depending on whether partnerships with infected partners are overselected ($\alpha(t, \mu) < 1$) or underselected ($\alpha(t, \mu) > 1$) relative to partnerships with uninfected partners. This can be seen more clearly by assuming that $\alpha(t, \mu) \equiv \alpha$, and examining the relationship between the infection hazard function $h^*(t | \mu)$ associated with $S^*(t | \mu)$ and the "true" infection hazard function $\mu\lambda(t)$ associated with $S(t | \mu)$ in (5). From Appendix B:

$$h^*(t | \mu) = \lambda(t) \left[\frac{\mu}{1 - S(t | \mu)(1 - \alpha)} \right]. \quad (15)$$

Thus, for fixed contact rate μ the estimated hazard under selection bias is proportional to the underlying infectivity but is distorted to appear larger or smaller depending on whether $\alpha > 1$ or $\alpha < 1$, respectively. For $\alpha = 1$, $h^*(t | \mu)$ and $\mu\lambda(t)$ coincide. This relationship is illustrated in Figure 4 for the case of constant infectivity $\lambda = 0.01$, and for two choices of α . The figure and equations (14) and (15) also illustrate that the assumption of proportional hazards no longer holds when the data are subject to selection bias. From the figure, the estimated infection hazard when partnerships with infected partners are twice as likely to be selected as partnerships with uninfected partners ($\alpha = 1/2$) is distorted to appear twice as large initially, with the discrepancy with the true infectivity decreasing with longer duration of exposure.

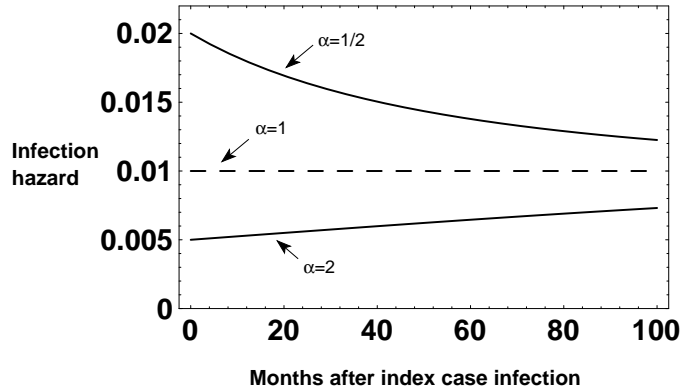


Figure 4. Infection hazards when partnerships with infected partners are overselected by a factor of two ($\alpha=1/2$) relative to partnerships with uninfected partners (top curve), and when the reverse selection frequency holds (bottom curve). Calculated using equation (15) assuming that the true infectivity is constant ($\lambda=0.01$) (middle line).

The impact of selection bias on the coefficient γ for the log contact rate $\log \mu$ in model (9) in the case where both infectivity and the ratio of selection probabilities are constant (and for fixed μ and t) is illustrated in Figure 5. (Appendix B gives details on the calculation of this bias.) Recall that when the true transmission model is given by (5) then $\gamma = 1$. It is clear from Figure 5 that selection bias can either attenuate or inflate the value $\hat{\gamma}$, depending on whether $\alpha < 1$ or $\alpha > 1$. As discussed at the beginning of this section, the situation where $\alpha < 1$, corresponding to higher selection probabilities for partnerships with infected partners is more likely to occur in practice. Thus, we would expect selection bias to result in attenuation of regression coefficients. Notice however, from Figure 5 that this attenuation should never be enough to change the sign of an estimated coefficient, even if selection probabilities are extremely discrepant. (This observation is verified in Appendix B.)

There are plausible reasons to expect that selection probabilities will depend on both duration of exposure and on other covariates describing exposure of partner to index case. A simple example of how this might arise is the case where couples with infected partners *and* shorter exposure are oversampled. This may occur in situations where couples with shorter exposure times (e.g. short-term monogamous relationships) have different reasons for enrolling into a study couples with longer exposure durations. For example, short-term partnerships may be more likely to enroll if partners who have been infected already know their serostatus. Even if the true infectivity is constant, this selection process will lead to the mistaken impression that the infection hazard is time dependent. In addition, as shown in Appendix B, estimated regression coefficients in transmission models such as (10) or (11) can actually differ in sign with their true values. This may lead to the mistaken impression that a covariate is associated with

decreased transmission risk when the opposite is true. (Note that this will occur even if more conventional techniques such as logistic regression are applied.)

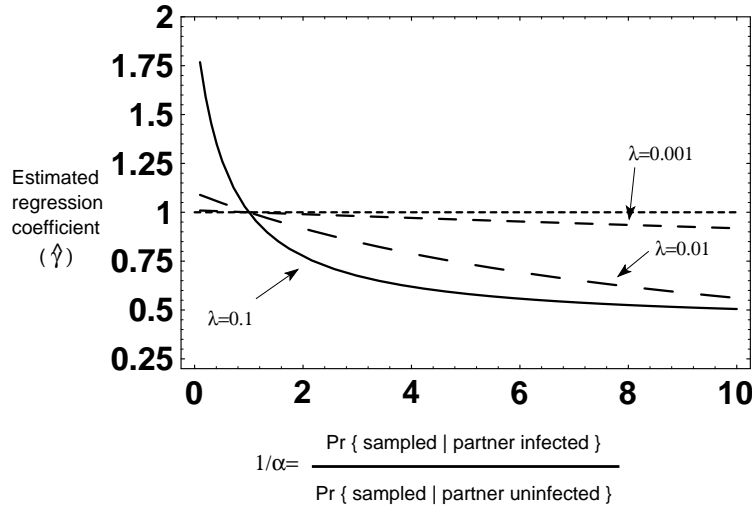


Figure 5. Attenuation in the estimated value of the regression parameter γ from model (9) due to biased selection of partnerships, shown as a function of the infectivity λ and the parameter α describing the degree of selection bias. The small dashed line represents no selection bias.

4.3 Measurement Error in Exposure Information

As discussed in Section 2, in most partner studies there will be concern over the accuracy of measurement of the exposure variables t and μ . Errors may arise because these variables are retrospectively ascertained, or in situations where the infection date of the index case is unknown.

Uncertainty about the date of index case infection corresponds to uncertainty about the duration of exposure t , and therefore about the contact rate μ applying to this period as well. This is illustrated in Figure 1. If a partnership has existed for a substantial period of time (Figure 1(b)), it may be reasonable to assume that the index case infection occurred after partnership formation. However, if the location of this event within the interval between the start of the relationship and recruitment into the study is unknown, the duration of exposure t is also known only to be of no greater length than this interval. For partnerships of relatively brief duration (Figure 1(a)), unless it can be established that the index case was infected before initiation of contact (or unless the partnership is mutually monogamous), exposure information is also subject to measurement error. Thus, distinguishing between the two basic data structures pictured in Figure 1 may be difficult in practice. One common method for controlling for unknown time of index case infection is to impute this time as the earliest possible date infection

could have occurred, using knowledge about the development of the AIDS epidemic in the region. (For example, in San Francisco no infections were reported prior to 1978.) This procedure will lead to systematic overestimation of the duration of exposure for long-term partnerships (Figure 1(b)), and to dislocation of the exposure interval relative to the true time of index case infection in short-term partnerships (Figure 1(a)).

Even in cases where the index infection dates are known (e.g. transfusion associated infections), error in estimated contact rates (μ) can be expected; partnerships may alter their contact pattern in response to a number of factors, including knowledge of the HIV status and disease stage of the index case. Many partner studies attempt to control this by obtaining estimated rates before and after the partner became aware of the infection in the index case.

Statistical approaches to investigating the impact of measurement error usually are based on the assumption that errors are random rather than systematic. Jewell and Shiboski [1, 4] present an approach for assessing the impact of random measurement error in contact rates and counts on regression parameter estimates from the constant infectivity model. The remainder of this section deals with systematic measurement error introduced into the duration of exposure t via lack of knowledge of the date of index case infection, as introduced above.

4.3.1 Measurement Error in Duration of Exposure

As discussed at the beginning of this section, frequently no reliable date of infection is available for index cases in a retrospective partner study, and it is known only to lie in a broad interval (e.g. between the beginning of the epidemic and the time of enrollment). One common method for addressing this problem involves assuming that any contacts occurring within this interval are potentially exposed, which is equivalent to assuming that all index cases were infected at the beginning of the interval. Using notation introduced in Sections 2 and 3 and below, Figure 6 illustrates this situation for a long-term partnership. Define

- A = first possible chronological time of index case infection,
- B = chronological time of recruitment,
- I = chronological time of index case infection (unobserved),
- J = chronological time of partner infection (unobserved),
- T = assumed duration of exposure ($T = B - A$).

For long-term partnerships in which contact began before the imputed time A , Figure 6 indicates clearly that the estimated duration of exposure T is an overestimate of the actual duration $B - I$. Thus, basing inferences about properties of the distribution of the true time V between index case infection and partner infection on T rather than $B - I$ can be expected to lead to biased estimates of transmission probabilities, infectivity and regression coefficients. The rest of this section investigates the impact of this bias on estimation of the infectivity and regression coefficients using the methods described

in Section 3. Although similar issues apply to the case of short-term partnerships, the discussion is limited to long-term partnerships.

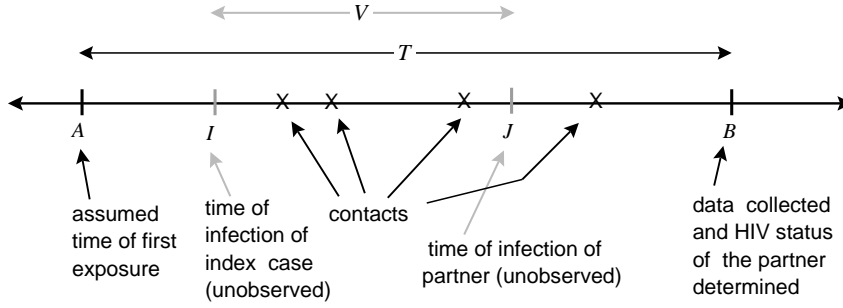


Figure 6. Illustration of exposure data from a long-term partnership for which the chronological time I of index case infection is unknown, and is estimated to occur at chronological time A . The assumed duration of exposure T is an overestimate of the true duration $B-I$.

Recently, Jewell, *et al.* [17] have described the conditions under which nonparametric estimation of the distribution of V is possible from data of the form described above, and have proposed an algorithm for estimation. Currently these methods do not provide for covariates and require some assumptions to be made about the form of the distribution of index case infections in chronological time, so they can not be used for infectivity estimation or for inference about the impact of risk factors on transmission. However, the results provide tools for investigating the bias issues raised in the previous paragraph.

Following Jewell, *et al.* [17], let G denote the distribution of the chronological time of infection for the index cases in a sample of n partnerships. (Note that this distribution may also vary between index cases due to differing knowledge about the period of infection.) The probability that the partner is observed to be infected at chronological time B , given only that the index case was infected sometime in the interval $[A, B]$ is given by the expression below:

$$\begin{aligned}
 \text{pr}(Y = 1 | A, B; A \leq I \leq B) &= \int_0^{B-A} \frac{G(B-u)}{G(B)} dF(u) \\
 &= \int_0^{B-A} \frac{F(u)g(B-u)}{G(B)} du \\
 &= \int_A^B \frac{F(B-s)}{G(B)} dG(s) \\
 &= \int_A^B \frac{F[T-(s-A)]}{G(B)} dG(s) ,
 \end{aligned} \tag{16}$$

where the second equality follows by an integration by parts, the third by the change of variable $s = B - u$ and the fourth from the correspondence $B = A + T$. (Note that $G(A)$

is assumed to equal zero in the above derivation, corresponding to the assumption that the index case could not have been infected before A .) Denote the probability defined in (16) by $F^*(T)$. Thus, for a given assumed duration of exposure t , $F^*(t)$ gives the probability that the partner is observed to be infected at B within t time units of A . The complementary probability associated with F^* is given by:

$$\begin{aligned} S^*(t) &= \text{pr}(Y = 0 \mid A, B; A \leq I \leq B) \\ &= \int_A^{A+t} S[t - (s - A)] g(s \mid A + t) ds, \end{aligned} \quad (17)$$

where $g(s \mid A + t) = g(s)/G(A + t)$, the conditional distribution of the chronological time of index case infection I given the presumed duration of exposure T and chronological start of exposure A . Note that $S^*(t)$ can be viewed as the survival function associated with the time between the assumed time of index case infection A and the time of transmission of the virus to the partner. This time, which is denoted by V^* , is related to V by the expression: $V^* = V + I - A$. Equation (17) elucidates the relationship between the distributions of V and V^* and provides a way to evaluate the biases which arise from analysis using the incorrect time scale. As shown below, there are similarities with bias issues which arise in the analysis of prevalent cohort data, as discussed by Brookmeyer and Gail [18].

First, the relationship between the induced survival function $S^*(t)$ and the true infection survival function $S(t)$ from (17) is addressed without considering covariate effects. (i.e. $S(t)$ in equation (17) can be interpreted as the baseline survival function given in equation (4) with $\mu \equiv 1$.) In Appendix C, it is established that in the case that the actual time until partner infection V follows an exponential distribution with constant infectivity λ , that the probability $S^*(t)$ of being infection free after a duration of exposure t measured on the incorrect scale will be systematically larger than the actual probability $S(t)$. Figure 6 makes this clear: if duration of exposure is systematically overestimated, then transmissions will be estimated to occur later than they actually do, giving the impression of lower transmission rates at any given time.

To investigate the impact of measurement error in exposure duration on covariate effects in proportional hazards models, consider a single covariate z in model (11) (z could also represent $\log \mu$ in (10)). If proportional hazards is assumed to hold for data based on the incorrect exposure duration t (i.e. model (11) is assumed to apply, when in fact (17) is correct), then the regression coefficient associated with z can be obtained from the relationship

$$\beta^* = \log[-\log S^*(t \mid \mu, z + 1)] - \log[-\log S^*(t \mid \mu, z)] .$$

This equation expresses the estimated regression coefficient β^* as a function of the true regression coefficient β . Appendix C investigates this relationship and provides some approximate results on the degree of bias expected in estimates of regression coefficients such as β^* . If the conditional distribution g is assumed not to depend on z , the following results are established: (i) if $\beta = 0$ then β^* will also be zero, and (ii) if $\beta \neq 0$, the value of β^* will be attenuated from β , where the degree of attenuation depends on

properties of the index case infection distribution g . As shown in the Appendix, even if there is a large amount of uncertainty about the true time of index case infection (e.g. this time is assumed to be uniformly distributed across the presumed exposure interval), the attenuation in β^* will be small relative to the magnitude of attenuation observed for uncontrolled heterogeneity in Section 4.1 (see Figure 3). Jewell and Shiboski [1] made a similar observation regarding the relative amount of attenuation resulting from random measurement error in contact information in the context of the constant infectivity model (1).

Note that the conditional distribution g in (17), that describes the distribution of index case infections in chronological time depends on the assumed duration of exposure t . If this distribution depends on the covariate z as well (i.e. if the time of index case infection is influenced by covariates considered in the analysis), the effect of z on transmission risk as measured by the coefficient β^* will be *confounded* with effects of z on exposure duration. This is analogous to the *onset confounding* described by Brookmeyer and Gail [18], but differs in the following ways: In the prevalent cohort problem, the initiating event (HIV infection) is unobserved but the final event (AIDS diagnosis) is observed. By contrast, in a retrospective partner study of long-term partnerships, both initiating event (index case infection) and terminating event (partner infection) are unobserved. Figure 6 illustrates that all that is known about the times of these two events (I and J) is that they lie in the assumed interval of exposure $[A, B]$, and therefore that $A \leq V \leq B$. As shown by Brookmeyer and Gail [18], no reliable inferences about effects of covariates in the framework of the proportional hazards model are possible when this type of confounding is present. In the case of a single covariate z , this means that the sign of β^* may differ from the sign of β , causing misleading conclusions about the effect z on transmission risk. (An example of this is provided in Appendix C.) Stratification of analyses by variables thought to be related to index case infection time is one method of controlling for this type of confounding. Examples of such variables include mode of infection of the index case and geographic region.

5 Examples

In this section, some of the methods described above are applied to current data from the California Partners' Study (CPS) [19, 20] and to data from the CDC Transfusion study [21] of heterosexual transmission. Data from 98 couples consisting of long-term female partners of male index cases from the CPS, and 55 long-term female partners of male index cases from the CDC study are considered. All partnerships considered satisfy the eligibility criteria outlined in Section 2, and by self-report, are monogamous (on the part of both partner and index case for the CDC study, and at least on the part of the partners for the CPS).

Partnerships in the CPS were recruited by self-referral, and by referrals from physicians, research studies, and local departments of public health. Index cases in this study were infected from a variety of sources, including blood transfusions, IV drug use and bisexual contacts. Dates of infection for the index cases are unknown, or are known

only to lie within a broad time interval. These dates of infection were estimated based on interview data, chronological time of enrollment and knowledge of the epidemic curve of HIV infections in California. Because these partnerships are long-term, it can be assumed that contacts began before the index case was infected (see Figure 1b for an illustration). For 10 index cases infected via blood transfusions, infections were assumed to occur on January 1, 1980. For 48 index cases infected through bisexual contact, infections were assumed to occur on January 1, 1982. For 24 index cases infected through IV drug use or through heterosexual contact, infections were assumed to occur on January 1, 1984. Sixteen index cases with no identified source of infection risk were also assumed to be infected on January 1, 1984. These dates were chosen to represent the left endpoint of the interval within which index cases were most likely to have been infected, based on the factors cited above. Several other schemes for assignment of index case infection dates were experimented with and did not influence the results reported here substantially. Finally, contact counts were calculated from reported contact rates μ according to interview information relative to the period of the partnership after the assumed date of infection of the index case.

Partnerships in the CDC study were recruited by identifying an AIDS diagnosis in the index case. Index cases in this study were all infected from blood transfusions. Thus, times of infections are known fairly accurately, and contact rates and counts can be calculated based on the actual duration of exposure. The median age of partners in the CPS and CDC study is 37 and 54 respectively.

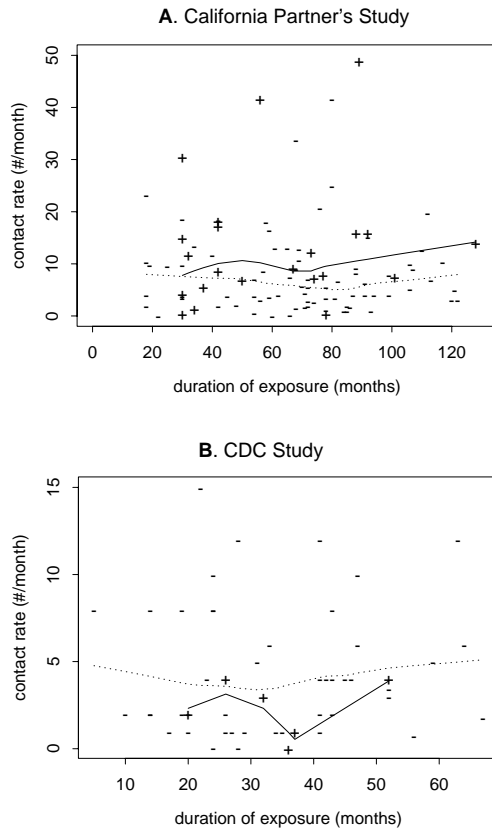


Figure 7. Observed contact rates (y-axes) plotted against duration of exposure (x-axes) for two partner studies. Partnerships with infected and uninfected partners are denoted by the symbols "+" and "-" respectively. The plotted lines refer to locally weighted regression smooths of the observed points for infected partners (solid lines) and uninfected partners (dashed lines).

Figure 7 displays observed contact rates (μ) plotted against durations of exposure (t) for the CPS and CDC data. The lines plotted in the figure represent the approximate mean contact rate expressed as a function of the duration of exposure. (These lines are locally weighted regression fits of the data [22].) Separate lines are shown for partnerships with infected partners and those with uninfected partners. Because of the retrospective nature of the sampling, we would expect that in a representative sample the mean contact rate would not change systematically with exposure duration. This translates to no apparent trends in the lines displayed in Figure 7. Although no strong trends are evident for either study, infected partners in the CPS report (on the average) slightly higher rates of contact than uninfected partners for almost all durations of

exposure. Conversely, in the CDC data, partnerships with infected partners report (on the average) lower contact rates than those with uninfected partners. Also, the range of contact rates reported in the CPS is much larger, possibly due in part to the relatively larger number of younger partners.

Table 1 presents a breakdown of the CPS exposure data by observed serostatus y and the estimated number of contacts $k = \mu \times t$, grouped into six intervals. This presentation emphasizes the structure of the data needed to fit the constant infectivity model (6).

Table 1. HIV seroprevalence from the California Partners' Study grouped by estimated numbers of contacts.

| <i>Serostatus of partner</i> | <i>Number of contacts ($\mu \times t$)</i> | | | | | |
|----------------------------------|---|--------|---------|---------|---------|----------|
| | 2-30 | 31-135 | 136-347 | 347-541 | 542-876 | 877-4355 |
| Seropositive (%) | 2 (40) | 2 (13) | 2 (11) | 4 (20) | 5 (26) | 7 (35) |
| Seronegative | 3 | 13 | 17 | 16 | 14 | 13 |
| Totals | 5 | 15 | 19 | 20 | 19 | 20 |

Table 2 displays results of fitting the constant infectivity model (6), the parametric alternative (9), and the semiparametric time dependent infectivity models (5) and (10) to data from both the CPS and CDC study. Goodness of fit of models in the table is assessed by examining changes in the deviance (related to the numerical value of the negative log-likelihood at its maximum for each model [5]) across models. The infectivity estimates obtained from the intercept terms in model (6) were 0.0004 and 0.0016 for the CPS and CDC data, respectively. These results are in the range of many published infectivity estimates. As discussed in Section 3.3.1, the goodness of fit of the constant model can be assessed by fitting the parametric alternative (9). This model provides a clear improvement over the constant model for both studies, and the attenuation from unity apparent in the estimated regression coefficients $\hat{\gamma}$ raises the possibility of time dependence in transmission risk, heterogeneity of infectivity, measurement error in exposure information or selection bias. The test of overdispersion described in Section 4.1 and Appendix A results in significant p values (at the 0.05 level) for both studies, indicating the possible presence of uncontrolled heterogeneity.

Table 2. Results from fitting transmission models (5), (6), (9) and (10) to data from two partner studies.

| Study | Fitted Model | Deviance | Change in | | γ | <i>p</i> (fit) | <i>p</i> (overdispersion) |
|--------------|---------------------|-----------------|--------------------|----------------------------|----------------------------|-----------------------|----------------------------------|
| | | | deviance | γ | | | |
| CPS | constant (6) | 114.60 | – | | 1.00 | – | 0.02 [♦] |
| | parametric (9) | 102.72 | 11.88 [*] | | 0.23 | 0.001 | |
| | time dep. (5) | 100.37 | 14.23 [*] | | 1.00 | 0.08 [◇] | |
| | time dep. (10) | 94.55 | 5.82 [▲] | | 0.40 | 0.02 | |
| CDC | constant (6) | 52.95 | – | | 1.00 | – | 0.004 [♦] |
| | parametric (9) | 35.47 | 17.48 [*] | | –0.31 | <0.001 | |
| | time dep. (5) | 49.41 | 3.54 [*] | | 1.00 | 0.3 [◇] | |
| | time dep. (10) | 32.83 | 16.58 [▲] | | –0.37 | <0.001 | |

* change in deviance from the constant infectivity model (6)
 ▲ change in deviance from the time dependent infectivity model (5)
 ♦ approximate *p* values, obtained from score test for overdispersion described in Appendix A.
 ◇ approximate *p* values comparing fit of time dependent model (5) to the constant model (6), obtained from a simulation test based on 200 replications.

The improvement in fit of the semiparametric time dependent model (5) over the constant model was assessed using a simulation test based on 200 replications (described in [2]). For the CPS, this model shows a modest improvement in fit over the constant model. For the CDC study, this is not the case. However, model (10), which generalizes (5) by allowing the value of the coefficient γ to differ from one, fits significantly better than (5) for both datasets. The estimates $\hat{\gamma}$ from these models display attenuation to a similar degree as in model (10), indicating that even when possible time dependence is accounted for, factors such as heterogeneity of infectivity, measurement error in exposure information and selection bias may be present. As shown in Section 4, these factors may all lead to apparent time dependence in transmission risk where none exists.

The negative estimate of γ in models (9) and (10) for the CDC data seems to indicate a decrease in cumulative transmission risk with increasing duration of exposure and frequency of contact. This seemingly paradoxical result has led some investigators to suggest that transmission probability may not depend on exposure [23], or that transmission risk may fail to increase with increasing numbers of exposed contacts [11]. As shown in Section 4, some forms of biased selection of partnerships and/or heterogeneity of infectivity may also lead to this conclusion. Because only partnerships with index cases developing AIDS in the recruitment interval were enrolled into this study, the selection was biased to include index cases who progressed to AIDS faster than average. In fact, the median incubation time for this sample is 2.6 years. This time is considerably shorter than the median of approximately ten years reported from many other

sources [24]. If shorter incubation is related to likelihood of virus transmission, then this selection process combined with apparent heterogeneity could lead to the negative estimates in Table 2.

Figure 8 displays the fit of the constant infectivity model (6) to the California Partners' Study data. The empirical complementary log-log transformed values of the grouped proportions of seropositive partners in Table 2 are shown for reference. The fit of model (9) is also shown. Clearly, (6) fits poorly for small numbers of contacts, while (9) provides a better overall fit. This observation is confirmed by the results in the Table 2. Note that the results in Section 4 indicate that infectivity estimates (based on the intercepts in the figure) from either of these models may be biased by the factors discussed above.

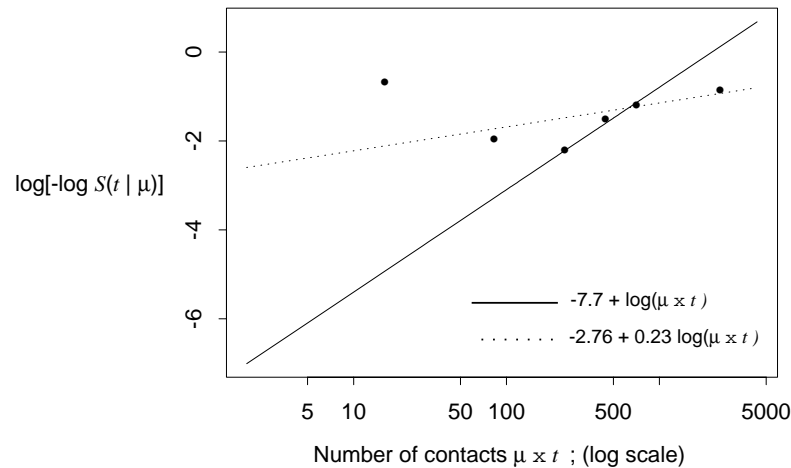


Figure 8. Fitted lines from the constant infectivity model (6) (solid line) and the alternative parametric model (9) (dashed line) for the California Partners' Study data. Plotted points refer to the complementary log-log transformed proportions of infected partners from Table 2.

Figure 9 displays estimated cumulative infectivities and infectivities from fitting model (5) to the data from the CPS and CDC studies. Infectivity estimates were obtained from the log cumulative infectivities in (5) by differencing and smoothing. The results are remarkably similar from both studies, with apparent increases soon immediately after index case infection, and rapid decreases following approximately two years of exposure. As discussed in Section 4, these decreases are consistent with the hypothesis of neglected heterogeneity of infectivity. (Recall that evidence of heterogeneity was found for both studies in the results reported above.) Therefore, both the scale and shape of the estimates in Figure 9 must be called into question.

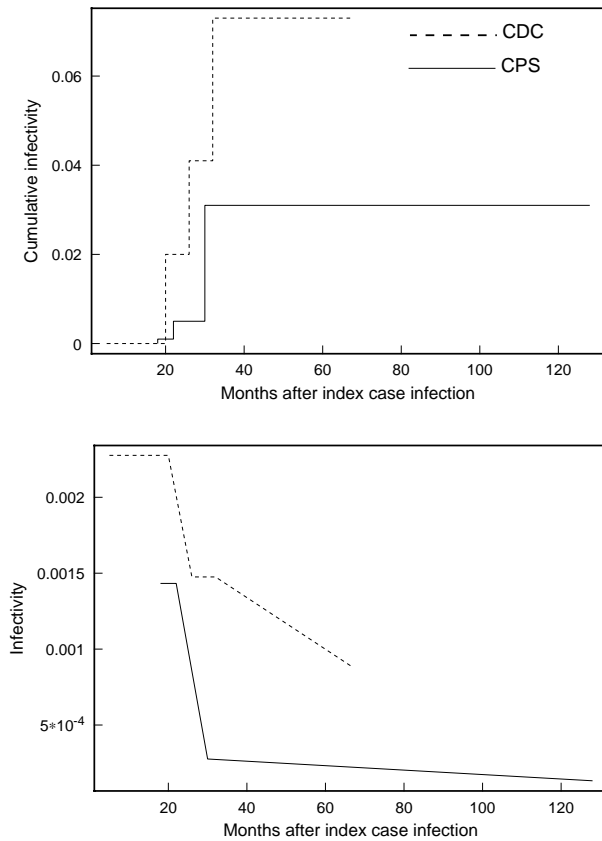


Figure 9. Cumulative infectivity (top) and infectivity (bottom) functions estimated by fitting model (5) to data from the CDC study (dashed line) and the California Partners' Study (CPS, solid line), plotted against length of exposure following time of index case infection .

6 Conclusions

The intent of this Chapter has been to provide a framework for statistical interpretation of data from partner studies of sexual transmission of HIV. The methods presented emphasize evaluating the plausibility of the assumption of constant infectivity, and attributing departures from this assumption to the influence of several factors, including time variation of infectivity, partnership heterogeneity in infectivity not accounted for by measured covariates, systematic and random measurement error in exposure variables and biased recruitment of partnerships. The influences of these factors are not unique to partner studies, and may operate in any investigation of HIV transmission. However, because of the retrospective nature of the data in most partner studies, the biases resulting from these factors may be particularly severe.

Although the results indicate that data from retrospective partner studies can frequently not be used to obtain accurate estimates of infectivity and transmission probabilities, several conditions are given which can guide interpretation of observed effects of risk factors from these studies. In general, heterogeneity, selection bias and measurement error in exposure data all can result in *attenuation* of the measured effects of risk factors from their true values. This attenuation is rarely extreme enough to change the sign of estimated regression coefficients. Thus, although the magnitude of the effect of a covariate cannot be accurately estimated when these sources of bias are operating, the direction of the effect will in general be estimated correctly. This means that partner studies can still yield valuable data on risk factors for HIV transmission. However, in extreme cases of bias, no reliable inferences can be drawn from such data, and observed effects of risk factors can differ in both sign and magnitude from their true values. One situation in which this may occur is when selection probabilities for enrollment differ for partnerships with infected and uninfected partners, and in addition, depend on other variables used in statistical analyses such as measures of exposure duration or intensity. Another example is provided by studies in which index case infection times are known only to lie in a broad interval, inducing systematic measurement error in estimated exposure duration and intensity. In this case, if the distribution of index case infection times depends on other variables considered in an analyses (analogous to *onset confounding* in prevalent cohort studies of HIV progression), estimated regression coefficients can differ in sign from their true values and misleading inferences may result.

The methods developed have been applied to data from two recent partner studies of heterosexual HIV transmission. Infectivity estimates from these studies are remarkably similar, and indicate possible time variation in infectivity with time following index case infection. However, diagnostic results also indicate that substantial partnership heterogeneity in infectivity may be present in data from both studies, which may lead to bias in shape and scale of the estimates. Previous analyses of the data from the CDC study have interpreted the lack of fit of the constant infectivity model as evidence for the hypothesis that transmission probability does not depend on exposure [11, 23, 25]. The results presented here indicate that the biased nature of the sample of index cases recruited into this study, as well as possible heterogeneity of infectivity noted by other authors [26] provide equally plausible explanations.

A number of important topics were not addressed in this chapter. Although bias in parameter estimates from a variety of sources is discussed, bias in corresponding variance estimates is not treated. Understanding of this is necessary for a complete picture of how inferences are affected by the sources of bias investigated here. In addition, the results focus on partner studies of transmission in monogamous relationships with known index cases. Understanding how exposure and other factors influence transmission risk in situations where the source of infection is unknown, or where multiple sources of infection is a possibility is much more difficult. Magder and Brookmeyer [27] present an approach to analyzing data from partner studies where the index case is unknown. Finally, the applicability of the results to sexually transmitted diseases other than HIV is not discussed. Because similar difficulties arise in studying transmission of

other sexually transmitted diseases, it is likely that many of the results reported here are relevant in design and analysis in studies of other diseases as well. In addition, since most of the results are based on the retrospective or *current status* nature of partner study data, they should apply to other studies based on this observation scheme.

Acknowledgements

The author would like to thank Dr. Nancy Padian and Dr. Thomas Peterman for permission to use data from their studies, and the reviewer for suggestions and comments which were very helpful in revising the manuscript. This research was supported by an award from CFAR (Center For AIDS Research), University of California, San Francisco, via a grant from the National Institute of Allergy and Infectious Diseases (grant # P30 AI 27663), and by SIMS (Societal Institute for the Mathematical Sciences) via a grant from the National Institute on Drug Abuse (grant # DA04722).

References

- [1] Jewell NP, Shiboski SC. Statistical analysis of HIV infectivity based on partner studies. *Biometrics* 1990;**46**:1133-1150.
- [2] Shiboski SC, Jewell NP. Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association* 1992;**87**:360-372.
- [3] Kim MY, Lagakos SW. Estimating the infectivity of HIV from partner studies. *Annals of Epidemiology* 1990;**1**:117-128.
- [4] Jewell, NP, Shiboski SC. The design and analysis of partner studies of HIV transmission. In: Ostrow DG, Kessler R, eds. *Methodological Issues in AIDS Behavioral Research*. New York: Plenum Publishing Company, 1993.
- [5] McCullagh P, Nelder JA. *Generalized linear models*, 2nd ed. New York: Chapman and Hall, 1989.
- [6] Cox DR, Oakes D. *Analysis of survival data*. New York: Chapman and Hall, 1984.
- [7] Hastie TJ, Tibshirani RJ. *Generalized additive models*. New York: Chapman and Hall, 1990.
- [8] Bacchetti P. Additive isotonic models. *Journal of the American Statistical Association* 1989;**84**:289-294.
- [9] Aalen OO. Heterogeneity in survival analysis. *Statistics in Medicine* 1988;**7**:1121-1137.
- [10] Diamond ID, McDonald JW. The Analysis of current status data. In: Trussel J, Tilton J, eds. *Demographic Applications of Event History Analysis*. Oxford: Oxford University Press, 1991.
- [11] Kaplan EH. Modeling HIV infectivity: must sex acts be counted? *Journal of Acquired Immune Deficiency Syndromes* 1990;**3**:55-61.
- [12] Ho DD, Moudgil T, Alam M. Quantitation of Human Immunodeficiency Virus type I in the blood of infected persons. *New England Journal of Medicine* 1989;**B321**:1621-1625.
- [13] Groeneboom P, Wellner JA. *Information bounds and nonparametric maximum likelihood estimation*. DMV Seminar series, Band 19. Basel: Birkhäuser Verlag, 1992.
- [14] Hougaard P. Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika* 1984;**71**:75-83.

- [15] Cox DR. Some remarks on overdispersion. *Biometrika* 1983;**70**:269-274.
- [16] Lancaster, T. *The Econometric analysis of transition data*. New York: Cambridge University Press, 1990.
- [17] Jewell NP, Malani HM, Vittinghoff E. Nonparametric estimation for a form of doubly censored data with application to two problems in AIDS. (Submitted for publication in *Journal of the American Statistical Association*.)
- [18] Brookmeyer R, Gail MH. Biases in prevalent cohorts. *Biometrics* 1987;**43**:739-749.
- [19] Padian N, Marquis L, Francis DP, Anderson RE, Rutherford GW, O'Malley PM, Winkelstein W. Male-to-female transmission of Human Immunodeficiency Virus. *Journal of the American Medical Association* 1987;**258**:788-790.
- [20] Padian N, Shiboski S, Jewell NP. The association between number of exposures and rates of heterosexual transmission of Human Immunodeficiency Virus (HIV). *Journal of Infectious Diseases* 1990;**161**:883-887.
- [21] Peterman TA, Stoneburner RL, Allen JR, Jaffe HW, Curran JW. Risk of HIV transmission from heterosexual adults with transfusion-associated infections. *Journal of the American Medical Association* 1988;**259**:55-63.
- [22] Cleveland WS. Robust Locally weighted regression and smoothing. *Journal of the American Statistical Association* 1979;**74**:829-836.
- [23] Anderson RM, May RM. Epidemiological parameters of HIV transmission. *Nature* 1988;**333**:514-519.
- [24] Bacchetti P, Moss AR. Incubation period of AIDS in San Francisco. *Nature*, 1989;**338**:251-253.
- [25] Blower SM. Re: "A method for estimating HIV transmission rates among female sex partners of male intravenous drug users" (letter to the editor). *American Journal of Epidemiology* 1993;**136**:1532-1534
- [26] Wiley JA, Herschkorn SJ, Padian NS. Heterogeneity in the probability of HIV transmission per sexual contact: the case of male-to-female transmission in penile-vaginal intercourse, *Statistics in Medicine* 1989;**8**:93-102.
- [27] Magder L, Brookmeyer R. The analysis of infectious disease data from partner studies with unknown source of infection. *Biometrics* 1993;**49** (to appear).
- [28] Cox DR, Hinkley DV. *Theoretical statistics*. New York: Chapman and Hall, 1974.

A Investigating the impact of heterogeneity

A.1 Impact on estimates of infectivity and covariate effects

Assume V is a random variable representing a failure time, with conditional hazard function $h(v|w) = wh(v)$, where w is the observed value of an absolutely continuous nonnegative random variable W with unit expectation, distribution function $K(\cdot)$, and density $k(w)$. This is a case of the random hazard model introduced in Section 3.1.3. Following Hougaard [14] and Lancaster ([16], Chapter 4), the marginal survival function for V can be written

$$\bar{S}(v) = \int_0^\infty \exp\left[-\int_0^v wh(x)dx\right] dK(w) = L(M(v)), \quad (18)$$

the Laplace transform of the random variable W with respect to the integrated hazard function $M(v) = \int_0^v h(x)dx$. The hazard function associated with this survival function can be written

$$\begin{aligned} \bar{h}(v) &= -\frac{d}{dv} \log \bar{S}(v) = \frac{L'(M(v))}{L(M(v))} \frac{d}{dv} M(v) \\ &= h(v) \frac{L'(M(v))}{L(M(v))} = h(v) E[W | V \geq v], \end{aligned} \quad (19)$$

The last equality follows from the fact that

$$\frac{L'(M(v))}{L(M(v))} = \int_0^\infty w \left[\frac{\exp\{-\int_0^v wh(x)dx\} k(w)}{L(M(v))} \right] dw, \quad (20)$$

where the bracketed term represents the conditional distribution of W given that $V \geq v$. Substituting the hazard function $\mu\lambda(v)$ into (19) and (20) for $h(v)$ verifies equation (12). From (19),

$$\frac{d}{dv} \log \bar{h}(v) = \frac{d}{dv} \log h(v) + \frac{d}{dv} \log E[W | V \geq v].$$

Using the identity for $E[W | V \geq v]$ in (20), the last term in the preceding equation can be shown to be strictly negative, i.e.

$$\frac{d}{dv} \log E[W | V \geq v] = -h(v) \frac{\text{Var}[W | V \geq v]}{E[W | V \geq v]}. \quad (21)$$

With the substitution $h(v) = \mu\lambda(v)$, this verifies equation (13).

Next, the impact of heterogeneity on covariate effects is considered. Assume that the "true" model is of the form defined at the beginning of this appendix, with conditional hazard function $wh(\cdot)\exp(\gamma z)$, with W defined as above, z the observed value of a fixed covariate Z and regression coefficient γ (usually assumed to be one for $z = \log \mu$). As in Section 3.1.3, assume that W is independent of Z . The appropriate survival function to use when W is unobservable is then given by (18). If the heterogeneity introduced by W is ignored, and the proportional hazards model assumed to apply, then

the fitted regression coefficient γ^* can be expressed as a function of the true survival function (and hence as a function of the true regression coefficient γ) as follows:

$$\gamma^* = \frac{d}{dz} \log \bar{h}(v | z).$$

The function defined in this equation can be computed from (19) using a calculation similar to that leading to (21) (see [16]):

$$\begin{aligned} \gamma^* &= \gamma + \frac{d}{dz} \log E[W | V \geq v] \\ &= \gamma \left\{ 1 - M(v | z) \frac{\text{Var}[W | V \geq v]}{E[W | V \geq v]} \right\}, \end{aligned} \quad (22)$$

where $M(v | z) = \int_0^v h(x) \exp(\gamma z) dx$. This result has two implications for interpreting models fitted to data with unobserved heterogeneity: (i) as a result of the heterogeneity introduced by W , the marginal distribution of the data is no longer of the proportional hazards form; and (ii) the fitted regression coefficient will be attenuated from (but will have the same sign as) the true value, with the attenuation given by the multiplicative factor in (22).

Implicit in the results of the previous paragraph is the assumption that the distribution $k(w)$ of the mixing variable W does not depend on the covariate Z (see Section 3.1.3). If this is not the case, then a calculation similar to that leading to (22), but based on the conditional distribution of W given z (i.e. using the density $k(w | z)$), can be used to show that the estimated coefficient γ^* equals the expression given in (22), plus the additive term given below:

$$\frac{\int_0^\infty S(v | w, z) k'(w | z) dw}{\int_0^\infty S(v | w, z) k(w | z) dw} \left(\frac{E[W | V \geq v, z]}{1 - E[W | V \geq v, z]} \right),$$

where $S(v | w, z)$ refers to the conditional survival function (with hazard function $h(v)w \exp(\gamma z)$), and k' denotes the derivative of the conditional density of W with respect to z . The sign of this term depends on properties of the conditional density function, and on the magnitude of $E[W | V \geq v, z]$ (recall that $E[W] = 1$). For certain choices of these factors γ^* may differ in both sign and magnitude from γ .

A.2 A score test to detect heterogeneity

Following a proposal made by Cox [15], consider an approximation to the marginal survival function $\bar{S}(v)$ (defined in (18)), obtained by expanding the conditional survival function $S(v | W) = \exp[-WM(v)]$ in a second order Taylor series about the mean of the mixing random variable W (assumed to be one):

$$S(v | W) \approx \exp[-M(v)] \left\{ 1 - M(v)(W - 1) + [M(v)(W - 1)]^2 / 2 \right\}.$$

Taking expectations with respect to W yields

$$\bar{S}(v) \approx \exp[-M(v)] \left\{ 1 + \frac{\sigma^2}{2} [M(v)]^2 \right\}, \quad (23)$$

where $\varepsilon = \sigma^2/2$, and σ^2 represents the variance of W . The right hand side of this equation still defines a survival function [16]. If the integrated hazard function $M(v)$ is given a parametric form (e.g. constant hazard), this expression can be used as the basis for a score test of the hypothesis $H : \sigma^2 = 0$ of no heterogeneity. Recalling the binomial form of the likelihood introduced in Section 3.2, standard results from likelihood theory [28], lead to the score statistic U^2/V_e . The components of this statistic are defined below, where L denotes the logarithm of the likelihood function (8) based on the marginal survival function (23), $\bar{F}(v) = 1 - \bar{S}(v)$, and the integrated hazard function $M(v) = \lambda v$:

$$U = \left[\frac{\partial}{\partial \varepsilon} L(\lambda, \varepsilon) \right]_{\lambda=\hat{\lambda}, \varepsilon=0} = \sum_{i=1}^n \left(\frac{y_i - \bar{F}(v_i)}{\bar{F}(v_i) (1 - \bar{F}(v_i))} \right) \frac{\partial \bar{F}(v_i)}{\partial \varepsilon},$$

and $V_\varepsilon = \mathfrak{t}_{\varepsilon\varepsilon} - \frac{\mathfrak{t}_{\varepsilon\lambda}^2}{\mathfrak{t}_{\lambda\lambda}}$.

Here, $\mathfrak{t}_{\varepsilon\varepsilon}$, $\mathfrak{t}_{\varepsilon\lambda}$ and $\mathfrak{t}_{\lambda\lambda}$ are elements of the expected information matrix evaluated at the maximum likelihood estimate $\hat{\lambda}$ obtained under the null hypothesis of no heterogeneity. These functions are defined individually as follows:

$$\mathfrak{t}_{ee} = \left[\sum_{i=1}^n \frac{(\partial \bar{F}(v_i)/\partial \varepsilon)^2}{\bar{F}(v_i) (1 - \bar{F}(v_i))} \right]_{\lambda=\hat{\lambda}, \varepsilon=0}, \quad \mathfrak{t}_{e\lambda} = \left[\sum_{i=1}^n \frac{(\partial \bar{F}(v_i)/\partial \varepsilon) (\partial \bar{F}(v_i)/\partial \lambda)}{\bar{F}(v_i) (1 - \bar{F}(v_i))} \right]_{\lambda=\hat{\lambda}, \varepsilon=0},$$

$$\mathfrak{t}_{\lambda\lambda} = \left[\sum_{i=1}^n \frac{(\partial \bar{F}(v_i)/\partial \lambda)^2}{\bar{F}(v_i) (1 - \bar{F}(v_i))} \right]_{\lambda=\hat{\lambda}, \varepsilon=0}.$$

Based on standard results from likelihood theory [28], the statistic U^2/V_e follows a χ^2 distribution with one degree of freedom. The development above can be easily generalized to account for more than one parameter in the integrated hazard function $M(v)$.

B Evaluating the impact of selection bias

Let W be a random variable representing selection of a partnership into a sample:

$$W = \begin{cases} 1 & \text{if partnership is selected} \\ 0 & \text{otherwise} . \end{cases}$$

Recall that Y denotes the binary response observed on each partner. Let t denote the observed length of exposure and \mathbf{z} be a vector of observed covariates (which may include the log contact rate $\log\mu$), define

$$\begin{aligned} \pi_1(t, \mathbf{z}) &= \text{pr}(W = 1 \mid Y = 1, t, \mathbf{z}), \text{ and} \\ \pi_0(t, \mathbf{z}) &= \text{pr}(W = 1 \mid Y = 0, t, \mathbf{z}) , \end{aligned}$$

the selection probabilities conditional on the observed value of the response, exposure time and the covariates. Then a standard probability argument for binary responses based on retrospectively sampled data [5] leads to the following expression for the response probability for the sampled observations:

$$\text{pr}(Y = 1 \mid W = 1, t, \mathbf{z}) = \frac{\text{pr}(Y = 1 \mid t, \mathbf{z})}{\text{pr}(Y = 1 \mid t, \mathbf{z}) + \frac{\pi_0(t, \mathbf{z})}{\pi_1(t, \mathbf{z})} \text{pr}(Y = 0 \mid t, \mathbf{z})}$$

Recalling definition (2) from Section 3.1.2, and setting $\pi_0(t, \mathbf{z}) = \alpha(t, \mathbf{z})\pi_1(t, \mathbf{z})$, this can be re-expressed as

$$1 - S^*(t \mid \mathbf{z}) = \frac{1 - S(t \mid \mathbf{z})}{1 - S(t \mid \mathbf{z}) + \alpha(t, \mathbf{z})S(t \mid \mathbf{z})} ,$$

where $S^*(t \mid \mathbf{z})$ denotes the survival function for the sampled data. Suppressing the dependence on \mathbf{z} , the associated hazard function can be written as

$$h^*(t) = h(t) \left\{ \frac{1}{1 - S(t)[1 - \alpha(t)]} \right\} - \frac{\alpha'(t)}{\alpha(t)} \left\{ \frac{1 - S(t)}{1 - S(t)[1 - \alpha(t)]} \right\} . \quad (24)$$

For the case of equal selection probabilities ($\alpha(t) \equiv 1$) the hazard function $h^*(t)$ for the observed data and the true hazard are identical. For $\alpha(t) \equiv \alpha$, indicating no time dependence in the selection process, the second term above drops out leaving

$$h^*(t) = h(t) \left[\frac{1}{1 - S(t)(1 - \alpha)} \right] . \quad (25)$$

At $t = 0$, $h^*(0) = h(0)/\alpha$. Thus, when the selection probability for an infected partner ($Y = 1$) is greater(less) than that for a uninfected partner ($Y = 0$), the hazard for the observed data is initially distorted to appear greater(less) than the true hazard. As t increases, $h^*(t)$ approaches the true hazard. For the extreme case $\alpha(t) \simeq 0$, corresponding to selection probabilities for infected partners being uniformly much greater than

for uninfected partners, the observed hazard is approximately equal to $h(t)/[1 - S(t)]$, which is infinite for $t = 0$, and decreases to $h(t)$ as t increases.

To examine the impact of selection bias on regression coefficient estimates, assume that a single covariate z modifies the true baseline hazard $h(t)$ via a proportional hazards assumption, so that $h(t | z) = h(t) \exp(\beta z)$. Let β^* denote the regression coefficient obtained from fitting (5) to the data (i.e. proportional hazards is incorrectly assumed to hold). The relationship between β^* and β is obtained from the expression

$$\beta^* = \frac{d}{dz} \log h^*(t | z)$$

which is a complicated function of β . Denote this function by $Q(\beta)$. First, assume $\alpha(t, z) = \alpha$. In this case, using (25), and expanding the right-hand side of the previous equation in a first order Taylor series about $\beta = 0$ gives the result that $\beta^* \cong \beta Q'(0)$, where

$$Q'(0) = \left[\frac{\partial}{\partial \beta} Q(\beta) \right]_{\beta=0} = \frac{S(t)(1 - \alpha)[1 + h(t)] + \alpha}{S(t)(1 - \alpha) + \alpha}, \quad (26)$$

and $S(t)$ denotes the true baseline survival function (i.e. $S(t | 0)$). Substituting in the chosen form for $S(t)$, $h(t)$ and α , this equation can be used to compute the approximate multiplicative bias in regression coefficients from models fitted to data subject to the form of selection bias described above (see Figure 5 for an example). Note that for $\alpha \in [0, 1]$, corresponding to attenuation in the estimated value of β , $Q'(0) > 0$ in (26). Thus, although selection bias may attenuate the estimated β^* from the true value β , the sign of β^* will match that of β .

For the more general case, in which $\alpha(t, z)$ is allowed to depend on t and/or z , the attenuation factor can be calculated similarly (using (24)), leading to a complicated expression for β^* of the form:

$$\beta^* \approx Q(0) + \beta Q'(0). \quad (27)$$

Here, $Q(0)$ and $Q'(0)$ both may depend on t , z and properties of $S(t)$. In this case, β^* and β may actually differ in sign. As an illustration, assume $\alpha(t, z) = \exp(\eta z)$. Then (25) applies, with α replaced by $\exp(\eta z)$, and both $Q(0)$ and $Q'(0)$ in (27) depend on z through $\exp(\eta z)$. Assume z takes on only the values zero and one. At $z = 0$, calculations show that to first order

$$\beta^* = \beta - \eta S(t),$$

thus the sign of β^* may be different from that of β , depending on the direction and degree of selection effect as determined by η .

C Measurement error in exposure data

Recalling definitions introduced in Section 4.3.1, let T denote the presumed duration of exposure, measured from chronological time A to recruitment at time $B = A + T$. Let $g(\cdot | t)$ be the conditional density of the chronological time I of index case infection given that this time is known to lie in the interval $[A, A + T]$. For convenience, assume that $A \equiv 0$, which is the same as assuming that exposure began simultaneously for all partnerships in a sample (e.g. A could correspond to January 1, 1978). Then, the probability that a partner is observed to be infection-free following an exposure of presumed length not greater than T (defined in equation (17)) is given by

$$\begin{aligned} S^*(t | z) &= \int_0^t S(t-s | z) g(s | t) ds \\ &= \int_0^t \exp \left[- \int_0^{t-s} \lambda(u) e^{\beta z} du \right] g(s | t) ds, \end{aligned} \quad (28)$$

where $g(s | t) = g(s)/G(t)$, z denotes the observed value of a covariate (e.g. the contact rate μ) and β is the associated regression coefficient. Given that the chronological time of index case infection is s , $S(t-s | z)$ represents the true probability of observing a partner to be infection-free following the actual duration of exposure $t-s$. Thus (28) represents the probability of the partner remaining infection-free when the duration of exposure and the time of index case infection (corresponding to the initiation of exposure) are known only to lie in the interval $[0, t]$. As discussed in Section 4.3.1, this probability can be regarded as a survival function for the random variable V^* which measures time to partner infection on the incorrect time scale measured from the assumed beginning of exposure at $A = 0$. Suppressing the dependence of $S^*(t | z)$ on the covariate z , the hazard function corresponding to S^* is defined below:

$$\begin{aligned} h^*(t) &= - \frac{d}{dt} \log S^*(t) = \\ &= \frac{\int_0^t f(t-s) g(s | t) ds}{\int_0^t S(t-s) g(s | t) ds} - g(t | t) \left[\frac{1 - S^*(t)}{S^*(t)} \right]. \end{aligned} \quad (29)$$

In the case where the true hazard is constant ($h(t) = \lambda$) the first term in (29) reduces to λ , and

$$h^*(t) = \lambda - g(t | t) \left[\frac{1 - S^*(t)}{S^*(t)} \right]. \quad (30)$$

Equation (30) shows that if the true hazard is constant, the observed hazard will appear to vary with exposure duration and will be smaller than the true hazard. The nature of this variation is governed by properties of the conditional distribution $g(t | t)$ of the time of index case infection I given the assumed duration of exposure t . If the mass of the distribution G is concentrated on one time point, then it follows from (30) that $h^*(t) = \lambda$. In addition, it can be shown from (30) that, for non-zero and finite values of t , $h^*(t) \leq h(t)$. For example, if the random variable I is uniformly distributed over

$[0, t]$,

$$h^*(t) = \lambda - \frac{\lambda - F(t)/t}{F(t)},$$

where $F(t) = 1 - e^{-\lambda t}$. This approaches $\lambda/2$ in the limit as t decreases to zero, and decreases to zero as t increases to infinity. Thus the hazard function is systematically underestimated (and the survival function systematically overestimated) when the exposure duration is measured with error. The behavior of $h^*(t)$ when the underlying hazard $h(t)$ is non-constant is more complex. Special cases can be examined by substituting the chosen form of $h(t)$ into (29).

The relationship between the covariate z and the probability $S^*(t | z)$ in (28) is clearly not of the proportional hazards form. If proportional hazards is mistakenly assumed to hold, and a model of the form (11) is fit when (28) actually applies, the estimated coefficient can be obtained from the relation below:

$$\beta^* = \log[-\log S^*(t | z + 1)] - \log[-\log S^*(t | z)].$$

This equation expresses the estimate β^* as a function of the true regression coefficient β in (28). It follows from (28) that if $\beta = 0$ then $\beta^* = 0$ as well. To investigate the magnitude of β^* for non-zero values of β , recall the approach taken in Appendix B, and denote the function in the above equation by $Q(\beta)$. Expanding the right-hand side of the previous equation in a first order Taylor series about $\beta = 0$ gives the result that $\beta^* \approx Q(0) + \beta Q'(0)$, where $Q(0) = 0$,

$$Q'(0) = \frac{\int_0^t S(t-s) \log S(t-s) g(s | t) ds}{\left[\int_0^t S(t-s) g(s | t) ds \right] \log \left[\int_0^t S(t-s) g(s | t) ds \right]}, \quad (31)$$

and $S(t-s) = S(t-s | 0)$. Thus, $Q'(0)$ gives the approximate multiplicative bias in the coefficient β^* . Letting $\Psi = S(t-s)$, the numerator in (31) can be written $E[\Psi \log \Psi]$, where the expectation is taken with respect to the conditional distribution of the unknown infection time I given the presumed duration of exposure t . The function $-\Psi \log \Psi$ is concave over $[0, 1]$, so Jensen's inequality for conditional expectations implies that $E[-\Psi \log \Psi] \leq (E\Psi)(-E \log \Psi)$. A second application of Jensen's inequality to the concave function $-\log \Psi$ shows that $E[-\Psi \log \Psi] \leq -(E\Psi)(\log E\Psi)$. It follows that, $|\beta^*| \leq |\beta|$, with equality holding only if the mass of the conditional infection distribution $g(s | t)$ falls on one time point (corresponding to all index cases becoming infected at the same time within the presumed exposure interval). In the special case where this distribution is uniform (i.e. $g(s | t) = 1/t$) and the underlying infectivity λ is assumed to be constant, the attenuation factor (31) can be easily calculated. It can be shown that even for relatively large values of λ and t the attenuation will be quite small. For example, assume model (6) holds (i.e. the true value of the regression coefficient associated with $\log \mu$ is unity) with $\lambda = 0.01$, and model (9) is fit. Then the observed value of γ will be no less than 0.9, even for extremely long exposure intervals.

When the distribution G of chronological time of index case infection depends on z as well as t , the relationship between the regression parameter β^* from fitting (28) and

the true regression coefficient β is more complicated. In this case, the approximation $\beta^* \approx Q(0) + \beta Q'(0)$ developed in the previous paragraph can be used to show that β^* can be larger, smaller or of different sign than β . To illustrate this, consider the special case where Z is a binary covariate taking on the values 0 or 1. Replacing the conditional density $g(s | t)$ in (28) with $g(s | t, z)$,

$$Q(0) = \log \left[\frac{\log \int_0^t S(t-s) g(s | t, 1) ds}{\log \int_0^t S(t-s) g(s | t, 0) ds} \right].$$

This term is zero when the distribution of index case infection times among partnerships with $z = 1$ is the same as those with $z = 0$ (i.e. when $g(s | t, 1) = g(s | t, 0)$), and ranges between positive and negative infinity otherwise. The term $Q'(0)$ is unchanged from (31). Although approximate, these results indicate that β^* can differ in both sign and magnitude from β .