The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 957-966

An Evaluation of the Accuracy of Maximum Likelihood Procedure for Estimating HIV Infectivity

Yonghwan Um[1] and Michael J. Haber[2]

Abstract

We evaluate the accuracy and precision of maximum likelihood estimation procedures for infectivity of HIV in partner studies. This is achieved by applying the procedure to hypothetical samples generated by computer. One hundred samples were generated with various combinations of parameters. The estimation procedure was found to be quite accurate. In addition, it was found that the power of the test for equality of infectivities for two types of contact depends on sample size and length of observation period, but not on the number of observations made on each subject. Tests based on a model for the infectivity had higher power than standard methods for comparing proportions.

1. Introduction

In order to formulate effective control strategies for the spread of the Human Immunodeficiency Virus (HIV) we need to understand the transmission dynamics and determinants of the epidemic. One of the major determinants is the infection transmission probability (infectivity). We define infectivity to be the probability of HIV transmission from a single sexual contact. Infectivity is an important measure used for evaluating the risks associated with particular sexual behaviours and for comparing the effectiveness of methods for controlling the HIV epidemic. Studies of the partners of the infected persons have provided estimates of infectivity. Several studies (Longini et al (1989), Wiley et al (1989), and Peterman (1986)) have found the per-contact male-to-female transmission probability to be about 0.001. Clark, et al. estimated the infectivity from a single unprotected receptive sexual contact with an infected partner to be 0.0027 among male homosexuals. Their estimates for anal and oral receptive contacts with non-steady partners were 0.0152 and 0.0041, respectively. DeGruttola, et al. (1989) reported an estimate of infectivity per receptive anal contact of about 0.005 to 0.010.

Clark, et al. developed a maximum likelihood (ML) procedure to estimate these

1) Assistant Professor, Department of Computer Science and Statistics, Sungkyul University, Anyang, Kyunggi-Do, Korea
2) Professor, Department of Biostatistics and Epidemiology, Emory University, Atlanta, GA, USA

probabilities. A number of studies show that the ML estimation technique gives good estimates. However, there are questions about the accuracy and precision of the procedure. A simulation model is used to evaluate the accuracy and precision of the ML procedure for estimating infectivities.

In this paper, firstly, we present a procedure for generation of random infection data when the number of partners and contacts, and the transmission probabilities are given. Data are generated using a range of parameter values and conditions. Secondly, from these simulated data, we use an ML technique, developed by Clark, et al, to estimate the infectivity of HIV associated with two types of sexual contacts which are referred to as anal and oral intercourse. Thirdly, by repeating the simulations under fixed conditions, we investigate the power of the test for the equality of the two infectivity parameters, the accuracy of the estimate of risk ratio of two parameters, and the coverage of the confidence interval for this risk ratio. We also investigate the effect of sample size, the length of the time between two observations on the same person and the number of observations on estimators and power.

2. Methods

2.1 Generation of infection data

In order to study the transmission of HIV, the computer generated hypothetical samples to which the simulation model was applied. The simulated samples were of size n = 200, 400, 800. Each study subject in the sample has one steady partner whose infection status (infected or not infected) is known and remains constant over time, as well as several non-steady partners with unknown infection status. Twenty five percent of partners (known and unknown) are infected. The n study subjects were assumed to be independent, i.e., the sample does not include the steady partner of a person in a sample. Each subject is followed for 10 equal time periods. The infection status of each subject can be determined at the end of each time period. The number of contacts of each type, per time period, with the known partner is uniformly distributed over {0, 4, 8, 16}, i.e., there is a 0.25 probability that 0 contacts occur, 0.25 probability that 4 contacts occur, etc. Similarly, the number of contacts with each unknown partner is uniformly distributed over {0, 2, 4, 8}. The number of unknown partners is uniformly distributed over {0, 1, 2, 4}.

There are two types of contacts with per-contact transmission probabilities, β_1 and β_2, where β_1 is the probability of transmission from an anal sexual contact with an infected partner, and β_2 is the probability of transmission from an oral sexual contact with an infected partner. Using all the information (the status of the known partner and the number of contacts of each type, the number of unknown partners and the number of contacts of each type with each unknown partner, and the probability that an unknown partner is

infected), we determine the probability that a particular susceptible person becomes infected during a given time period (see below). Then a random number between 0 and 1 is selected, and, if this number is less than the probability of infection, then that person is considered "infected". Once infected, the person is not considered in subsequent time periods. We assume that there are 2 level of infectiousness indexed by L = 0, 1, with L = 0 for an uninfected person, L = 1 for an infected. The follow up starts at time t(0), and the 10 observations on each person are made at times t(1), t(2), ···, t(10). We denote $\tau(K)$ the time interval [t(K-1), t(K)] (K = 1, 2, ···, 10), and we use S to index the types of sexual contacts with S = 1 for anal and S = 2 for oral contact.

In addition, we assume the following information is available at time t(K), (K = 1, 2, ···,10) for each person :

$$X(K) = \begin{cases} 1 \text{ if the person is infected by the time } t(K), \\ 0 \text{ otherwise} \end{cases}$$

L(K) = infectiousness level of the known partner of the study subject in time interval $\tau(K)$

D(K, S) = the number of type S contact with the known partner during time interval $\tau(K)$.

A(K) = the number of unknown partners of the study subject in the time interval $\tau(K)$.

B(K, S) = the average number of type S contacts with each unknown partner during time interval $\tau(K)$.

P(S, L) = probability of transmission in a type S contact with an infected partner at infectious level L.

R(L, K) = probability that an unknown partner is at the infectiousness level L at time t(K)

The primary parameters of interest use are transmission probabilities (infectivities) β_1 and β_2. We note that $P(1,1) = \beta_1$ and $P(2,1) = \beta_2$ and that P(1,0) = P(2,0) = 0. Considering an exposed person who was susceptible at time t(K-1), the probability that this person escapes infection from all contacts with the known infected partner during $\tau(K)$ period is given by :

$$Q_1(K) = \prod_S [1 - P(S, L)]^{D(K,S)} \quad (1)$$

We assume in eq. (1) that each contact is an independent event where HIV is transmitted with probability P(S,L). We further assume that the probability remains constant over the time periods. When making the same assumption, the probability that a person

escapes infection for all the contacts with unknown partners during $\tau(K)$ is

$$Q_2(K) = \left\{ \sum_L R(L,K) \prod_S (1 - P(S,L))^{B(K,S)} \right\}^{A(K)} \tag{2}$$

In our analysis, we assume that R(0,K) = 0.75 and R(1,K) = 0.25 for K=1,⋯,10, i.e., 25% of all potential partners are infected.

Using the assumption of independence of contacts with the known and unknown partners, the overall probability of escaping all the contacts with known and unknown partners during time interval $\tau(K)$ is

$$Q(K) = P(X(K) = 0 \mid X(K-1) = 0) = Q_1(K) * Q_2(K) \tag{3}$$

Therefore, the probability that a person who is susceptible at time t(K-1) becomes infected during the time interval $\tau(K)$ is 1-Q(K).

2.2 Estimation

The transmission probabilities β_1 and β_2 are estimated based upon the likelihood of observing a particular outcome sequence (X(1),⋯,X(10)) for each person in the cohort, given the information L(K), D(K,S), A(K), B(K,S) and R(L,K). Hence the probability that an exposed man becomes infected within period K is

$$P(X(1) = 0, \cdots, X(K-1) = 0, X(K) = 1) = \left[\prod_{m=1}^{K-1} Q(m) \right] (1 - Q(K)) \tag{4}$$

where K = 1, 2, ⋯, 10.

The probability that this person escapes infection for the entire period of observation is

$$P(X(1) = 0, \cdots, X(10) = 0) = \prod_{K=1}^{10} Q(K) \tag{5}$$

since we assumed that the distribution of X(K) given X(K-1) is the same for all K. The likelihood function for all the persons in the sample is

$$LF = \prod_{j=1}^{n} \left[\left[\prod_{m=1}^{K-1} Q(m) \right] (1 - Q(K)) \right]^{v_j} \left[\prod_{K=1}^{10} Q(K) \right]^{1 - v_j} \tag{6}$$

where v_j=1 if the person becoms infected during one of the K time intervals, v_j=0 otherwise for subject j (j = 1,⋯,n). The ML techniques is then applied to maximize the likelihood

function. This method provides estimates of P(S,L) and their covariance matrix (Mood et al 1974). Then the estimated values of P(S,L) can be compared to their preset values in the simulation program.

3. Results and Discussion

Table 1 shows the averages (over 100 simulations) of estimated β_1, β_2, β_1/β_2, the fraction of simulation where $H_0 : \beta_1 = \beta_2$ was rejected (power), the fraction of simulations where the confidence intervals (CI) for β_1, β_2, and β_1/β_2 included the true value, average values of the standard errors of the estimated β_1, β_2. For each simulation the 95% CI for β_1, β_2 and β_1/β_2 is given as $\hat{\beta}_i \pm 1.96\sqrt{\mathrm{var}(\hat{\beta}_i)}$, i=1, 2 and $(\exp(c_1), \exp(c_2))$, respectively where $c_1 \left(= \ln\frac{\hat{\beta}_1}{\hat{\beta}_2} - 1.96\sqrt{\mathrm{var}(\ln\frac{\hat{\beta}_1}{\hat{\beta}_2})} \right)$ and $c_2 \left(= \ln\frac{\hat{\beta}_1}{\hat{\beta}_2} + 1.96\sqrt{\mathrm{var}(\ln\frac{\hat{\beta}_1}{\hat{\beta}_2})} \right)$ are interval estimates of $\ln\frac{\beta_1}{\beta_2}$. And we used the statistic $Z = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\mathrm{var}(\hat{\beta}_1) + \mathrm{var}(\hat{\beta}_2) - 2\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2)}}$ for testing $H_0 : \beta_1 = \beta_2$ for each simulation. The estimated values $\hat{\beta}_1$, $\hat{\beta}_2$ are quite close to the preset values of β_1, β_2 yielding an error of no more than 0.001. The 95% CI for β_1, and β_2 contain the true value of β_1 and β_2 more than 95 times out of 100 simulations in most cases that observations were made at every time point. However, when there was only one observation (at time t(2)), the CI for β_1 gives 89% inclusion.

The values $\hat{\beta}_1 / \hat{\beta}_2$ give comparatively good estimates of β_1/β_2 for the case of $(\beta_1, \beta_2)=(0.01,0.01)$ but they are larger than true values in all other cases. The CI for β_1/β_2 contains the true values of β_1/β_2 about 95 times out of 100 simulations except for the case of $(\beta_1, \beta_2)=(0.04,0.005)$ when there were 10 observations. The power increases as β_1/β_2 increases in given sample size. This is obvious because β_1 gets different form β_2 as the ratio of β_1 and β_2 increases. The last 3 rows in Table 1 shows simulation results obtained when a single observation was made at the end of observation period. We see that for longer observation periods both the bias and standard error (S.E.) of the estimators β_1, β_2 decrease, while the power of the test for $H_0 : \beta_1 = \beta_2$ increases. The effect of sample size on estimators and power of the test for $H_0 : \beta_1 = \beta_2$ when observations were made at all 10 time periods is given Table 2. Increasing the sample size

has large effect on increasing the power. Table 3 shows the effects of the length of observation period and number of observations per subject when sample size is equal to 200. We see that the longer the observation period is, the larger the power is, but increasing the number of observations without changing the length of the observation period results in only a modest increase in power. For example the power increases from 0.30 to 0.86 as observation period increases from 2 to 10. However, there was only a small increase in power from 0.86, with a single observation at t(10), to 0.88, from 10 observations at times t(1),$\cdots$,t(10). This indicates that the power depends mainly on the time at which the last observation is made, but to a lesser extent on the number of observations. Table 4 shows the effect of the using an incorrect value of R(1,K) (the proportion of infected partners) in the estimation procedures for the case of n = 200, (β_1, β_2) = (0.01,0.005). There are large deviations of the estimated values $\widehat{\beta_1}$, $\widehat{\beta_2}$ from the true values of β_1, β_2 when the assumed proportion is not close to the true value of 0.25. The estimates of β_1 and β_2 decrease as the value used for R(1,K) increases. This indicates that not knowing the prevalence of HIV infection among partners may cause a large bias in the estimates of transmission probabilities. The estimated value of β_1/β_2 also increases as the estimate of R(1,K) increases, because $\widehat{\beta_2}$ changes in a faster rate than $\widehat{\beta_1}$ does. The power of the test for $\beta_1 = \beta_2$ tends to increase with the assumed value of R(1,K). Incorrect values of R(1,K) also affect the coverage of the confidence intervals. In order to compare the modeling approach with the standard method for comparing two risks, simulations were also carried out in a population of size n=200 where 100 subjects had only contacts of type 1 and the remaining subjects had only contacts of type 2 (see Table 5). A single observation was made after 10 time periods. The modeling approach gives higher power than the standard method because the model makes more efficient use of the contact structure of the data. This is consistent with the finding by Koopman, et al. (1989) that measures of risk based on transmission probabilities are more accurate and less subject to bias than those based on the usual odds ratios. We also see that while the standard method underestimates the risk ratio β_1/β_2, the method based on modeling overestimates that ratio.

4. Conclusions

The most important conclusions from this study are as follow:

(1) Modeling approach enables to estimate infectivities of two or more types simultaneously. It gives higher power than standard methods.

(2) Estimates of infectivities are usually quite accurate, but the method overestimates risk ratios.

(3) The length of the period of observation is important, but more intermediate observations are not important.

(4) It is important to have a quite accurate estimator of the proportion of infected partners.

Table 1. Simulations Results

n	times of observation	β_1	β_2	$\widehat{\beta_1}$	$\widehat{\beta_2}$	1000 x S.E. (β_1)	1000 x S.E. (β_2)	power*
200	1, 2, ···, 10	0.01	0.01	0.0099	0.0106	2.203	2.655	0.06
200	1, 2, ···, 10	0.01	0.005	0.0099	0.0052	2.021	1.922	0.28
200	1, 2, ···, 10	0.02	0.005	0.0201	0.0052	3.152	2.279	0.88
200	1, 2, ···, 10	0.04	0.005	0.0396	0.0056	5.243	2.755	0.99
400	1, 2, ···, 10	0.01	0.01	0.0099	0.0101	1.758	1.689	0.06
400	1, 2, ···, 10	0.01	0.005	0.0102	0.0051	1.578	1.212	0.49
400	1, 2, ···, 10	0.02	0.005	0.0203	0.0051	2.390	1.381	1.00
800	1, 2, ···, 10	0.01	0.005	0.0099	0.0051	1.079	0.860	0.84
200	10	0.02	0.005	0.0206	0.0051	3.780	2.362	0.86
200	5	0.02	0.005	0.0197	0.0053	4.025	3.060	0.67
200	2	0.02	0.005	0.0183	0.0075	5.362	5.004	0.30

* of the test for $H_0 : \beta_1 = \beta_2$

Table 1. continued

n	times of observation	β_1	β_2	coverage of CI for β_1	coverage of CI for β_2	β_1/β_2	estimate of β_1/β_2	coverage of CI for β_1/β_2
200	1, 2, ⋯, 10	0.01	0.01	96	95	1	1.04	0.96
200	1, 2, ⋯, 10	0.01	0.005	96	94	2	3.19	0.97
200	1, 2, ⋯, 10	0.02	0.005	98	97	4	4.81	0.95
200	1, 2, ⋯, 10	0.04	0.005	91	97	8	8.73	0.94
400	1, 2, ⋯, 10	0.01	0.01	94	95	1	1.05	0.95
400	1, 2, ⋯, 10	0.01	0.005	95	94	2	2.15	0.96
400	1, 2, ⋯, 10	0.02	0.005	98	95	4	4.41	0.95
800	1, 2, ⋯, 10	0.01	0.005	97	97	2	2.00	0.98
200	10	0.02	0.005	99	99	4	4.97	0.99
200	5	0.02	0.005	93	96	4	5.87	0.96
200	2	0.02	0.005	89	97	4	5.54	0.90

Table 2. The effect of sample size on power for testing $H_0 : \beta_1 = \beta_2$

n	power at given (β_1, β_2)		
	(0.01, 0.005)	(0.02, 0.005)	(0.04, 0.005)
200	0.28	0.88	0.99
400	0.49	1.00	—
800	0.84	—	—

Table 3. The effect on power of length of observation period and number of observation per subject for n= 200

Time points at which observation was made	power (β_1 = 0.02, β_2 = 0.005)
2	0.30
5	0.67
10	0.86
1, 2, ···, 10	0.88

Table 4. The effect of using an incorrect value of R(1,K) for n=200, (β_1, β_2)=(0.01,0.005), β_1/β_2=2. The true value is R(1,K)=0.25.

assumed value of R(1,K)	$\widehat{\beta_1}$	$\widehat{\beta_2}$	coverage of CI for β_1	coverage of CI for β_2	power*	estimate of β_1/β_2	coverage of CI for β_1/β_2
0.10	0.0144	0.0079	71	88	0.23	2.25	93
0.20	0.0114	0.0059	94	94	0.31	2.42	95
0.30	0.0090	0.0047	87	95	0.32	2.44	98
0.40	0.0075	0.0041	59	91	0.26	2.55	96
0.50	0.0062	0.0030	35	58	0.39	2.72	100

* of the test for $H_0 : \beta_1 = \beta_2$

Table 5. Comparison of the power of the test for $H_0 : \beta_1 = \beta_2$ based on the model with the standard method of comparing proportions.

Method	(β_1=0.01, β_2=0.005, β_1/β_2=2)		(β_1=0.02, β_2=0.005, β_1/β_2=4)	
	power	risk ratio	power	risk ratio
Model	0.46	2.42	0.96	4.32
Comparison of proportion	0.27	1.71	0.79	2.31

References

[1]. Longini IM, Clark WS, Haber M, et al. (1989). The stages of HIV infection: waiting times and infection transmission probabilities. In: Castillo-Chavez C. ed. *Lecture notes in biomathematics.* New York, NY: Spring-Verlag 83, 111-37.

[2]. Wiley JA. Herschkorn SJ. Padian NS. (1989).Heterogeneity in the probability of HIV transmission per sexual contact: the case of male-to-female transmission in penile-vaginal intercourse. *Statist Med* 8, 93-102.

[3]. Peterman TA, curran JW. (1986). Sexual transmission of human immunodefiency virus. *JAMA* 256, 2222-6.

[4]. Clark WS, Longini IM, Horsburgh CR, et al. the probability of HIV transmission through sexual exposures among male homosexual partners. (*unpublished manuscript).*

[5]. DeGruttola V, Seage GR, Mayer KH, et al. (1989). infectiousness of HIV between male homosexual partners, *J clin Epidemiol* 42, 849-56.

[6]. Mood AM, Grayvill FA, Boes DC, (1974). *Introduction to the Theory of Statistics.* 3rd ed. McGraw-Hill.

[7]. Koopman JS, Simon CP, Jacquez JA, et al. (1989). Seletive contact within structured mixing with an application to HIV transmission risk from oral and anal sex. In: *Lecture notes in biomathematics.* New York, NY: Springer-Verlag 83, 316-49.