Published in final edited form as:

Stat Med. 2010 December 10; 29(28): 2880–2889. doi:10.1002/sim.4062.

A modeling framework for the analysis of HPV incidence and persistence: a semi-parametric approach for clustered binary longitudinal data analysis

Xiangrong Kong^{a,b,*,†}, Ronald H. Gray^{a,c}, Lawrence H. Moulton^{b,c}, Maria Wawer^a, and Mei-Cheng Wang^b

^aDepartment of Population, Family and Reproductive Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, U.S.A.

^bDepartment of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, U.S.A.

^cDepartment of International Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, U.S.A.

Abstract

Human papillomavirus (HPV) infection is a common sexually transmitted disease of growing public health importance, and over 40 genotypes have been identified in genital infections. Current HPV cohort studies often follow participants at pre-determined visits, such as every 6-months, and data generated from such epidemiology studies can be described as clustered longitudinal binary data where correlation arises in two ways: the directionless clustering due to the multiple genotypes tested within an individual, and the temporal correlation among the repeated measurements on the same genotype along time. Current analyses for identification of risk factors associated with HPV incidence and persistence often either do not fully utilize information in the dataset or ignore the correlation between the multiple genotypes. Given the scientific definition of incidence and persistence, conditional probability modeling provides us a natural mathematical tool. We thus present a semi-parametric regression model for such data where full specification of the joint multivariate binary distribution is avoided by using conditioning argument to handle the temporal correlation and GEE to account for the correlation between the multiple genotypes. The model is applied to the HPV data from the Rakai male circumcision (MC) trial to evaluate the as-treated efficacy of MC and also identify modifiable risk factors for incidence and persistence of oncogenic HPV types in men. A simulation study is performed to provide empirical information on the number of individuals that is needed for satisfactory power and estimation accuracy of the association parameter estimates in future studies.

Keywords

clustered longitudinal binary data; conditional probability; semi-parametric model; HPV incidence; HPV persistence

1. Introduction

Genital human papillomavirus (HPV) infection is a sexually transmitted disease (STD) that is of growing public health importance as it is of high prevalence in both developing and developed countries, and is estimated to have the highest incidence of any STD in the United States [1]. The association between persistent high risk HPV infections and cancer also has been well established [1]. One unique feature of HPV is that based on HPV DNA homology, more than 40 genotypes have been identified to be related to genital HPV infections [2]. The different HPV types have different etiological characteristics and are associated with different clinical symptoms. Currently available HPV vaccines only have prophylactic efficacy on prevention against HPV type 6, 11, 16, and 18 in women between 15–26 years old. Other types are of increasing interest [3] with the concern that uptake of the HPV vaccines may lead to a shift of the distribution of HPV genotypes with increasing prevalence on non-vaccine targeted types. Therefore, it is important to investigate biological and epidemiological factors associated with type specific HPV incidence and persistence. Modern DNA testing technology allows the simultaneous detection and genotyping of multiple HPV genotype infections from a single DNA amplification assay, such as the commercial Roche HPV Linear Array (Roche Diagnostics, Indianapolis, IN) that can detect the presence or absence of 37 common types of HPV.

The genotype specific feature of HPV infections determines several statistical issues related to current HPV epidemiological studies. One is that given the sexually transmitted nature of HPV, the testing results on the multiple types within an individual are highly correlated. Also, in prospective studies participants are followed for a period of time during which genital samples are collected and tested periodically, yielding repeated measurements on the same genotype for each individual. Moreover, HPV infection is commonly a transient disease with an estimated median duration of 8 months for incident infections in women [1] and an estimated median duration of less than 6 months for prevalent infections in men [4,5]. Although it is not conclusive whether a preexisting infection that is undetectable by DNA assay has truly cleared or becomes latent, it is possible for an individual to re-acquire the same type of virus. Thus it can be observed in a longitudinal study that repeated events of acquisition and persistence of the same HPV genotype happen on the same individual. With these characteristics of data generated from longitudinal HPV studies, the data structure can be characterized as “*clustered binary longitudinal*” data, where correlation arises from two sources: the directionless clustering of the test results on the multiple types and the temporal correlation within test results on each specific genotype across study period.

Current HPV longitudinal studies often follow a cohort of individuals at pre-determined visits [6,7,8,9,10,11,12,13,14,15], such as every six months, and the published studies have mainly adopted two approaches to handle the complicated correlated HPV data. One is to use each individual as the unit of observation when studying HPV incidence or persistence (i.e. failure of clearance). Risk factors were then further identified using logistic or Poisson regression where individuals were classified as having or not acquired (or cleared) any HPV genotypes during the study interval [8,9,10,11,12]. For example, in [10,13], “multiple HPV types detected or cleared during a visit were treated as a single episode”. While it is essential to estimate HPV incidence among study participants, the data are not fully utilized when the scientific interest lies on identifying risk factors for HPV incidence or persistence. One can imagine that the disease burden for a person with one or two types of HPV infections is likely to be different from that for a person with multiple types infections (which is not uncommon in certain populations), yet they are treated in the model such that they make the same contributions to the estimation of the association parameters for the risk factors. Moreover, it is difficult in these models to incorporate time-

varying effects, such as sexual behaviors that are likely to change from time to time. An alternative approach is that each genotype was used as the unit of analysis and the log-rank test [4] or Cox regression model [14,15] was utilized to identify factors associated with HPV incidence or clearance. However, the correlation between the multiple coinfections within a person was often ignored, and the strength of the association of a risk factor with HPV incidence or persistence consequently would be inflated. Xue et. al. [16] recently provided a nice summary about the use of marginal and mixed effects models for analyzing HPV natural history data. For HPV incidence and persistence, they described the use of survival framework to model the time to first incident detection (or time to clearance), and Wei-Lin-Weissfeld method [17] in Cox regression models was used to account for the correlation between genotypes within person. The mixed effects alternative described is to use Cox regression model with a frailty effect to account for between-genotype correlation. However, in HPV longitudinal studies, genital samples are normally collected at pre-determined follow-up visits, such as every 6 months, thus the event time to incidence or persistence is actually a heavily interval censored survival outcome[18]. Also, given the disease characteristics that repeated incidence or persistence events can be observed on the same individual, only analyzing time to first incident detection does not make full usage of the entire longitudinal data. Analytical methods that can fully utilize the data, appropriately account for the correlations, and yield interpretable results are therefore needed for HPV epidemiology studies.

Statistical methods for analyzing general correlated binary data have been extensively studied, and can be summarized into a few classes including models based on the joint multivariate binary distribution [19], marginal models based on generalized estimating equation (GEE) approach [20,21], random effects models [22], and transitional (also called Markov, regressive, or conditional) models [23,24]. The best model to choose can only be determined with a clear understanding of the scientific question to be addressed. For the clustered binary longitudinal data observed in modern HPV epidemiology studies, there are two kinds of distinct correlations as discussed above. Our aim of this paper is to present a modeling framework for analyzing such HPV data when the scientific interest focuses on identification of risk factors for HPV incidence and persistence. The next section provides an example HPV dataset collected from a male circumcision (MC) randomized controlled trial conducted in Rakai, Uganda, and the data structure represents what is often observed in HPV epidemiology studies. In Section 3, we describe in detail our proposed analytical model, including the estimation procedure and asymptotic properties of the coefficient parameters. This model was applied to the MC trial data to evaluate the as-treated efficacy of MC in preventing HPV acquisition and persistence. Since knowledge of HPV infection in men is relatively limited compared to that for women, we also applied the model to identify demographic and behavioral factors that are associated with HPV incidence and persistence in men. The results of the applications are shown in Section 4. Borrowing the data from the MC trial, we conducted simulation to assess the power and estimate accuracy when the number of individuals is relatively small, and the simulation results are presented in Section 5. We then conclude the paper with discussions in Section 6.

2. An example of HPV epidemiology study-the Rakai male circumcision trial

A clinical trial of MC was conducted on initially HIV-negative uncircumcised men aged 15–49 years in Rakai District of Uganda during 2003–2006 [25]. Approximately 5000 men were enrolled and randomized to either immediate circumcision (intervention arm) or delayed circumcision after 24 months (control arm). At enrollment and follow-ups scheduled at 6, 12, and 24 months, penile swabs were collected and participants were interviewed to obtain information on sociodemographic characteristics, sexual risk behaviors and symptoms of sexually transmitted infections. The circumcision effect on HPV was a secondary trial

endpoint and HPV testing was performed restricted to consistently HIV-negative married men with concurrently enrolled wives [11]. Due to resource limitation, HPV testing was only conducted on samples collected at all four visits for a subset of randomly selected 330 such men (39.5%) in the intervention arm and 314 men (39.1%) in the control arm. Roche HPV Linear Array was used for HPV detection and genotyping of 37 common viral types. Therefore, for each participant at each visit, there are thirty seven binary indicators indicating the presence or absence of the HPV genotypes, respectively. As an example, Figure 1 shows the testing results on fourteen oncogenic high risk HPV types (HR-HPV) (HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, and 68) observed for one trial participant. Such repeated clustered HPV testing data together with the time varying exposure variables represent the data structure often encountered in HPV epidemiology studies. Following the definition used in the original publication of this study [11], HPV incidence was operationally defined as the acquisition of a new genotype during an interval (i.e. the individual was tested positive at the current visit but was negative for this HPV genotype at the prior visit). Persistence was defined as the failure of clearance of a pre-existing infection during an interval (i.e. the individual was positive at both the prior and the current visits).

3. Modeling framework for HPV incidence and persistence

3.1. Model formulation

Let $Y_{ij}(t)$ denote the binary outcome observed at the t th visit on HPV type j of individual i , where $t = 0, 1, 2, \dots, T$ and T is the number of follow-up visits; $j = 1, 2, \dots, J$ is the cluster size, i.e. the number of genotypes tested; and $i = 1, 2, \dots, n$ where n is the number of individuals in the study. Thus for individual i specifically, the vector $(y_{i1}(0), y_{i1}(1), \dots, y_{i1}(T); \dots, y_{iJ}(0), y_{iJ}(1), \dots, y_{iJ}(T))$ is the observed outcomes for all genotype at all visits. The joint probability function for participant i is (omitting the subscript i temporarily):

$$P(Y_1(0)=y_{1(0)}, \dots, Y_1(T)=y_{1(T)}; \dots, Y_J(0)=y_{J(0)}, \dots, Y_J(T)=y_{J(T)})=P(\mathbf{Y}(0)=\mathbf{y}_{(0)}, \mathbf{Y}(1)=\mathbf{y}_{(1)}, \dots, \mathbf{Y}(T)=\mathbf{y}_{(T)}) \quad (1)$$

where $\mathbf{Y}(t)$ is the $J \times 1$ vector for the outcome observed at time t .

It is important to notice that the scientific definition for “incidence/persistence” is in fact based on a conditional argument, hence conditional probability provides a natural mathematical modeling tool and can be used to account for the temporal correlation. Using the conditional probability property, the joint probability function in 1 can be expressed as:

$$\mathbf{Y}(0) \cdot \prod_{t=1}^T P(\mathbf{Y}(t)|\mathbf{Y}(t-1), \dots, \mathbf{Y}(1), \mathbf{Y}(0)) \quad (2)$$

Since the commonly used follow-up interval in HPV epidemiology studies is 6 months or longer, and HPV infection is only transiently detectable with a median duration of 6–8 months [1], it is therefore reasonable to assume that the current outcome only depends on the testing result observed at the prior visit, that is, we consider first order dependence: $P(\mathbf{Y}(t)|\mathbf{Y}(t-1), \dots, \mathbf{Y}(1), \mathbf{Y}(0)) = P(\mathbf{Y}(t)|\mathbf{Y}(t-1))$. Also, due to the genotype-specific feature, past infection of and immunity to one genotype generally does not provide cross-type protection against the future infection with other genotypes [26]. Thus we only care about the type-specific dependence, and there is

$$P(Y(t)|Y(t-1))=P(\{Y_1(t)|Y_1(t-1)\}, \dots, \{Y_j(t)|Y_j(t-1)\}, \dots, \{Y_j(t)|Y_j(t-1)\}) \quad (3)$$

To assess the association of exposure variables with HPV incidence and persistence, we use the following model which incorporates the dependence on the past result:

$$g(P(Y_j(t)=1|Y_j(t-1)=y_{j(t-1)}))=\mu+\alpha'\mathbf{x}_t+\beta y_{j(t-1)}+\gamma'\mathbf{x}_t y_{j(t-1)} \quad (4)$$

where $g(\cdot)$ is a link function, such as logistic link or log link that is often used for binary response, and $\mathbf{x}_t$ is a $P \times 1$ vector of exposure variables measured at the t th visit.

3.2. Interpretation of coefficient parameters

It is apparent that the event $\{Y_j(t) = 1|Y_j(t-1) = 0\}$ represents the operationally defined incidence, and $\{Y_j(t) = 1|Y_j(t-1) = 1\}$ represents the operationally defined persistence. Therefore, if using logistic regression in model 4, $\exp(\alpha_p)$ is the odds ratio for the occurrence of an incident infection when the exposure variable x_p ($p = 1, \dots, P$) increases one unit, while fixing the levels of other exposure variables; whereas $\exp(\alpha_p + \gamma_p)$ is the odds ratio for having a persistent infection during the interval for variable x_p . The intercept μ corresponds to the odds of incidence of a new infection and $\mu_0 + \beta$ corresponds to the odds of persistence of a pre-existing infection, respectively, for the baseline exposure group. Similarly, if using log link in 4, we can obtain incidence/persistence risk ratio estimates for the exposure variables.

3.3. Time heterogeneous model

Model 4 is essentially a time homogeneous first order transition regression model, and is applicable for most HPV epidemiologic studies where the length of follow-up intervals is fixed and the effects of risk factors are constant over the entire study period. For the HPV data from the Rakai circumcision trial, however, the follow-up interval during the trial varied from 6 to 12 months. In this situation, we can induce a time heterogeneous model by letting the association and baseline risk parameters be a function of the length of follow-up interval. Specifically, we let

$$\begin{cases} \mu=\mu_0+\mu_1 l_t \\ \alpha=\alpha_0+\alpha_1 l_t \\ \beta=\beta_0+\beta_1 l_t \\ \gamma=\gamma_0+\gamma_1 l_t \end{cases}$$

where $l_t = 0$ if the interval length is 6 months and $l_t = 1$ if the interval length is 12 months. Thus the time heterogeneous model is as:

$$g(P(Y_j(t)=1|Y_j(t-1)=y_{j(t-1)}))=\mu_0+\alpha_0'\mathbf{x}_t+\beta_0 y_{j(t-1)}+\gamma_0'\mathbf{x}_t y_{j(t-1)}+[\mu_1+\alpha_1'\mathbf{x}_t+\beta_1 y_{j(t-1)}+\gamma_1'\mathbf{x}_t y_{j(t-1)}]l_t \quad (5)$$

The time-heterogeneous model can also be applied in studies with long term follow-up where risk factors are expected to have time-varying effects on HPV incidence/persistence, such as studies of the long-term efficacy of HPV vaccines, where the indicator variable l_t in model 5 may be used to differentiate between short-term vs. long-term effect. The model may be further extended to accommodate other forms of time-varying effects.

3.4. Estimation of model parameters

From 3, the joint probability function of all individuals is

$$\prod_{i=1}^n P(Y_i(0)) \cdot \prod_{t=1}^T P(\{Y_{i1}(t)|Y_{i1}(t-1)\}, \dots, \{Y_{iJ}(t)|Y_{iJ}(t-1)\}) \quad (6)$$

Recall that the sexual transmission route of HPV determines that the incidence and persistence of each type is likely to be highly correlated with those of other types, but here this correlation is not of primary scientific interest, and the GEE technique [20] can be used to treat the nuisance correlation between the HPV types. That is, the estimation can start by maximizing the log “pseudo-likelihood” as if the conditional distributions of the multiple genotypes were independent:

$$L(\theta) \propto \sum_{i=1}^n \sum_{t=1}^T \sum_{j=1}^J [y_{ij}(t) \log p_{ij}(t) + (1 - y_{ij}(t)) \log(1 - p_{ij}(t))] \quad (7)$$

where $\theta = (\mu, \alpha, \beta, \gamma)$, $p_{ij}(t) = P(Y_{ij}(t) = 1 | Y_{ij}(t-1))$ is given in the regression model 4 or 5. Note that the marginal probability $P(Y(0))$ in 6 reflects the **prevalence** of type specific HPV infection at enrollment, and accordingly does not influence the estimation of parameters for incidence/persistence. To account for the correlation between multiple types, GEE theory[20], can be used to update the estimates for θ . That is, we seek the solution $\hat{\theta}$ for the generalized estimation equation of

$$\sum_{i=1}^n \sum_{t=1}^T \left(\frac{\partial \mathbf{p}_i(t)}{\partial \theta} \right)^T \text{Cov}^{-1}(\mathbf{Y}_i(t)) [\mathbf{Y}_i(t) - \mathbf{p}_i(t)] = 0 \quad (8)$$

where $\mathbf{p}_i(t) = (p_{ij}(t))$ is the $J \times 1$ vector of conditional means that relates to θ through model 4 or 5, and $\text{Cov}(\mathbf{Y}_i(t))$ is the $J \times J$ covariance matrix for the J outcomes observed at time t for individual i . Specifically, it is reasonable to use the exchangeable correlation matrix between different HPV types.

Since the temporal correlation within each genotype is taken care of by the conditional probabilities and between type correlation is treated with GEE, existing software such as SAS PROC GENMOD can be used to fit model 4 or 5. Within each individual, the rows for the same genotype across time can be viewed as independent, whereas at each time point, rows for different genotypes are correlated with the exchangeable correlation structure.

3.5. Another angle to look at the regression models

Model 4 or 5 is proposed starting from the joint probability function of the multivariate binary outcome. Another angle to look at the model is based on quasi-likelihood theory [27, 28]. Markov regression models for univariate time series were proposed in [29] where the modeling framework was derived from quasi-likelihood theory. Specifically for binary outcome, we can directly model the conditional mean for the random variable $Y_j(t)$, given $Y_j(t-1)$, as a function of the explanatory variables $\mathbf{x}$ and $y_j(t-1)$. Since $E(Y_j(t)|Y_j(t-1)) = P(Y_j(t) = 1 | Y_j(t-1))$, this model can be of the form in our earlier presented model 4 or 5. Moreover, since the conditional distribution is still a binary distribution, the variance of $Y_j(t)|Y_j(t-1)$ is a function of its mean, hence the model on the conditional mean falls in the generalized linear model framework [30]. Recall, however, that we have clustered binary longitudinal data from HPV studies, but again we can easily extend the univariate

conditional models by incorporating GEE adjustment on the covariance to account for the correlation between the multiple genotypes.

3.6. Asymptotic properties of the coefficient parameters

The asymptotic situation is the traditional scenario where the cluster size J is fixed, and the number of independent clusters $nT \rightarrow \infty$. For the binary conditional distribution and logistic linear model, the regularity conditions hold easily, and thus there is:

Theorem: $\hat{\theta}$ is asymptotic consistent and multivariate normally distributed, i.e.

$$\sqrt{N}(\hat{\theta} - \theta) \rightarrow N(\mathbf{0}, V) \text{ as } N \rightarrow \infty, \quad (9)$$

where $N = nT$ and V is the asymptotic covariance matrix given by:

$$\lim_{N \rightarrow \infty} N \left[\sum_{i=1}^n \sum_{t=1}^T E \frac{\partial \mathbf{U}_i(t)}{\partial \theta} \right]^{-1} \cdot \left[\sum_{i=1}^n \sum_{t=1}^T E(\mathbf{U}_i(t)^T \mathbf{U}_i(t)) \right] \cdot \left[\sum_{i=1}^n \sum_{t=1}^T E \frac{\partial \mathbf{U}_i(t)}{\partial \theta} \right]^{-1} \quad (10)$$

where $\mathbf{U}_i(t) = \frac{\partial \mathbf{p}_i(t)}{\partial \theta}^T \text{Cov}^{-1}(\mathbf{Y}_i(t))[\mathbf{Y}_i(t) - \mathbf{p}_i(t)]$ is the generalized score function as in 8, and $\hat{\theta}$ can be used to substitute θ to obtain a consistent estimate for the variance of $\hat{\theta}$.

The proof of the theorem follows from that for the asymptotic properties of GEE estimates in [20], with an extension that the independent sample size here is $n \times T$ as the conditional variable is independent across the T intervals, and thus is omitted. Note that for time series data, the asymptotic property of the quasi-likelihood model on the conditional mean in [29] depends on $T \rightarrow \infty$, whereas in HPV longitudinal studies, the number of follow-ups is often small. Nonetheless, the asymptotic property here relies on nT and often the number of individuals is quite large. This also shows that using conditional probability to treat the temporal correlation in HPV epidemiology studies not only yields desirable interpretation relevant to the scientific interest on incidence/persistence, but also increases the efficiency of the parameter estimates with T contributing to the number of independent samples.

4. Results for the Rakai circumcision trial

The effect of circumcision on HPV incidence and clearance (or persistence) was a secondary endpoint in the trial. At enrollment, men in the intervention and control arm were highly comparable with respect to sociodemographic and sexual behavioral characteristics [11], as well as the prevalence of HPV infection, with 36.7% men positive on at least one HR-HPV genotype in the intervention arm and 36.6% in the control arm, indicating a successful result of randomization. The follow-up was semiannual during the first year and annual in the second year. Since the oncogenic HR-HPV types are of particular public health importance, we only considered the fourteen HR-HPV genotypes and applied the presented modeling framework to evaluate the as-treated-efficacy of circumcision on HR-HPV incidence and persistence, and also to identify socialdemographic, general hygiene, and sexual behavioral factors that are associated with male HR-HPV incidence and persistence. The time-heterogeneous model 5 was first used to investigate whether the incidence risk (or odds) ratio or persistence risk (odds) ratio between two exposure groups differed during the 6-month interval compared to 12 month-interval. If they were not significantly different at level 0.05, then for model parsimony, the time-homogeneous model 4 was fit to obtain the risk or odds ratio estimates.

Table 1 shows the as-treated incidence/persistence risk ratios of the intervention arm compared to the control arm using the time-homogeneous model. Specifically we used the log link function in order to compare the results to those from the original paper reporting this dataset [11]. Clearly circumcision has significant efficacy against both HR-HPV incidence and persistence. The risk of having an occurrence of a new infection among the circumcised men is 45% less than that among the uncircumcised men (risk ratio=0.55, 95%CI: 0.42–0.73). In the original paper [11], the estimated incidence rate ratio of men having acquired any new HR-HPV infection was 0.67 (95%CI: 0.50–0.90). This underestimates the reduction of risk of new infections afforded by MC, as men having one infection were treated equally during the analysis to men having multiple infections. Table 1 shows that circumcised men also had 33% (two-sided $p=0.012$) lower risk of persistence of a pre-existing infection compared to the control arm men. Note that although the persistence risk is higher during the 6-month interval than that during the 12-month interval, the persistence risk ratio is not significantly different between the 6-month and 12-month intervals. Thus the time-homogeneous model is adequate.

Table 2 shows incidence/persistence odds ratios associated with the potential risk factors from univariate analyses. The logistic link was used to avoid convergence problems and possible invalid probability estimates when using the log link function with continuous predictor variables. The risk of HR-HPV acquisition in men declined with age as what has been well known in women [1]. The increased incidence with higher education was also observed with HIV incidence in the Rakai population [25], and possibly reflects higher rates of casual sex among the more affluent men with higher education. It is not surprising that having non-marital or non-consensual relationships or multiple partnerships significantly increased the risk of HR-HPV acquisition. The variable about sexual frequency (number of coital acts during the past 30 days) was treated as continuous and is significantly associated with HR-HPV incidence. Condom use was not protective and the utility of condoms use for HPV prevention has been controversial [1]. That inconsistent use of condoms significantly increased risk of HPV acquisition may be a reflection of relationships with non-marital partners. No variable is identified for HPV persistence, partly due to the smaller number of persistence events compared to the number of incidence events observed in the data.

A multivariate model with logistic link was also fit that included the as-treated status of circumcision and the covariates in Table 2. The adjusted incidence odds ratio of circumcision is 0.53 (95%CI: 0.40–0.71), and persistence odds ratio is 0.55 (95%CI: 0.36–0.82), indicating stable efficacy of circumcision in preventing penile HR-HPV infections and reducing risk of persistence of pre-existing infections. The significant inverse association between age and HPV incidence remains in the multivariate model, with odds ratio 0.67 (95%CI: 0.50–0.91) between age group of “31–40” and “ ≤ 30 ”, and 0.46 (95%CI: 0.27–0.78) between age group of “ > 40 ” and “ ≤ 30 ”. Among the behavioral variables, only number of sex partners has borderline significant influence on HPV incidence, with odds ratio 1.44 (95%CI: 0.98–2.12). Having multiple partners undoubtedly increases the risk of being exposed to HPV carried by female partners, and the result conforms to the sexually transmitted nature of the infection. The absence of statistical significance of sex frequency and condom use in the multivariate model implies that partner’s HPV carriage status is the fundamental determinant of HPV incidence observed in these men.

5. Simulation

In HPV epidemiology studies, often hundreds or thousands of individuals are followed for a few (2 or more) intervals, indicating that the sample size of independent observations nT is adequate for hypothesis testing and estimation based on the consistency and asymptotic

normality in 9. Note that the larger cluster size J the more independent observations are needed for the asymptotic normal distribution to hold. Here we further performed simulation to examine the power and estimation accuracy when the number of individuals is relatively small, as may be the situation in pilot studies. To mimic the correlation structure for HPV longitudinal data, we simulated data with smaller n using the observed data from the Rakai MC trial. That is, treating the repeated HPV testing results observed on the $n = 644$ men as the population, we randomly selected a smaller subset of individuals and assume their data are the sample at hand. This way of generating the simulation datasets has the advantage that the correlation between the repeated measurements on the same HPV type across time and the correlation between the incidence/persistence of multiple types are fully retained. Moreover, the magnitudes of HPV type specific prevalence, incidence and persistence, as well as the effect sizes of exposure variables, are meaningful as they are close to what we would observe in real world. Specifically, only HR-HPV data were generated, and the number of independent individuals simulated was $n = 50, 100, 150, 200, 300, 400$. Under each scenario, 500 simulation runs were performed.

With each simulated dataset, the time-homogeneous model 4 with the logistic link was used to estimate the incidence odds ratio (IOR) and persistence odds ratio (POR) associated with circumcision status. Note that the “true” effect size can be viewed as known from the whole observed data, with $\alpha = -0.60$ ($\log(\text{IOR})$) and $\alpha + \gamma = -0.54$ ($\log(\text{POR})$). Since hypothesis testing is important when the scientific interest focuses on risk factors identification, we estimated the power in testing the null hypotheses on incidence and persistence odds ratios, respectively, by the proportion of simulation runs rejecting H_0 (at level 0.05) among the total 500 runs. Estimation is also important once a significant variable has been identified. We evaluated the accuracy of estimates for IOR and POR, defined as the proportions of simulation runs where $\hat{\alpha} \in (-0.8855, -0.3238)$ and $\hat{\alpha} + \hat{\gamma} \in (-0.9456, -0.1276)$, respectively, among the total runs, where the parentheses are the 95% confidence intervals of the true effect sizes for IOR and POR from the whole observed data.

Figure 2 shows the power and estimation accuracy estimated from the 500 simulation runs, corresponding to testing/estimating IOR and POR, respectively. The number of available persistence events (i.e. $Y_t = 1 | Y_{t-1} = 1$) is generally smaller than the number of incidence events (i.e. $Y_t = 1 | Y_{t-1} = 0$), thus the power for testing the hypothesis about POR is always lower than that for testing IOR. Estimation accuracy is higher than the power both for IOR and POR, especially when n is small, implying that, as expected, the IOR and POR estimates have high variance when n is small. When $n = 200$, the power for testing IOR can achieve 74.0%, and the estimation accuracy is 80.4% for IOR and 77.8% for POR. However, the power for testing POR is only 29.2%. Therefore, identification of risk factors associated with persistence requires either a large number of individuals or a longer follow-up. Note the simulation study was based on simulation parameters derived from the Rakai MC study, with three follow-up visits and 35% prevalence of HR-HPV at enrollment. For studies with more follow-ups or populations with different levels of prevalence, other simulation parameters will be needed.

6. Discussion

Nowadays HPV infection is receiving growing public health attention with its high prevalence and incidence and its association with invasive cancers. Modern DNA technology provides an efficient platform for simultaneous detection and genotyping of multiple HPV types. The data generated from HPV epidemiology studies can be summarized as clustered longitudinal binary data where two distinct kinds of correlation exist: the correlation between the repeated measurements for each genotype over time, and the correlation between the different genotypes. Analysis of such data needs to take the

correlations into account appropriately. We presented a modeling framework for analyzing longitudinal HPV data when the scientific interest focuses on the identification of risk factors associated with HPV incidence and persistence. The joint modeling of incidence and persistence is essentially a “combo” treatment corresponding to the two correlations, which is motivated by the fact that incidence/persistence are conditional events. The model derivation determines that the coefficient parameters have desired interpretations for epidemiologists, and time-varying exposure variables can be naturally incorporated during the modeling, which are especially useful when examining behavioral variables. Moreover, since conditional probability is used to account for the temporal correlation, the independent sample size is multiplied by the number of follow-up intervals, yielding higher efficiency for parameter estimates. The correlation between genotypes is normally not of primary interest, nevertheless, the alternating logistic regression GEE model [31] can be used to obtain the odds ratio estimate which is a better measure for correlation between binary variables for HPV co-infections.

In the proposed models 4 and 5, since only the distribution of $Y_t|Y_{t-1}$ is specified through the regression model, the method is basically a semi-parametric approach. It can be viewed as a special case of the class of models for multivariate binary time series described in [32], where the model was induced from a log-linear model on the joint distribution of the multivariate binary data. Model 4 is a time-homogeneous model which implicitly assumes a constant effect of risk factors over the study period. The time-heterogeneous model is also presented for the situation where the length of study intervals varies, as in the Rakai MC data; or when risk factors may have time-varying effect on HPV incidence/persistence, as in studies with long follow-up. Following the operational definition of incidence/persistence used in the Rakai MC data, we assumed that the current HPV testing result solely depends on the testing result at the prior study visit, that is, only first order models were considered. This is reasonable given the known duration of HPV infection and for most HPV studies samples are collected semiannually. Nonetheless, the proposed models can be extended to accommodate higher order dependence, such as second order transition models that may be needed if the study interval is shorter, or when the operational definition of incidence and persistence relies on two prior results. We also assumed that for each HPV type, its current testing result does not depend on the infection history of other types. The biological difference between type-specific infections and lack of cross-genotype protections generally support this assumption[26]. However, one can statistically explore the validity of this assumption using the proposed modeling framework.

There are limitations to the presented modeling framework. With the operational definition of incidence commonly used in the HPV literature (either depending on one or two prior assay results), one cannot differentiate between an incidence infection and recurrent infection. Moreover, one cannot differentiate between reactivation of a previously latent infection or reinfection following a prior clearance event. This limitation however is determined by the current lab technology in that the assay cannot differentiate between the detection of a true new infection or recurrent infection. The latter scenario is more likely to concur with high risk sexual behaviors, and an incident infection observed in the current study period may still be a recurrent infection since the infection history before enrollment is unobserved. An analogous problem arises with clearance because one cannot differentiate between “true” clearance of infection and undetectable latent infection. Nevertheless, given the current assay technology, in studies with long follow-up where associations between risk factors and HPV incidence/persistence are expected to vary, the time-heterogeneous model 5 may be extended where a time variable can be used to indicate the stage of the study. The longer the study follow-up, the more likely that the individuals have been exposed to HPV infections and immunity may have developed during the course of the study. Thus the association parameters related to the time variable can reflect the change of risk for HPV

incidence/persistence over time. Another limitation of the presented method is that for persistence, the model models the risk of persistence of a pre-existing infection at the subsequent follow-up visit, hence it does not consider how long that infection has existed. When the scientific interest is the duration of infection and the study design has relatively short follow-up intervals, it is more intuitive to directly model the duration of persistence (i.e. time to clearance), and survival analysis especially multivariate survival models will be useful. In addition, missing data may occur in HPV studies, especially in studies on men. Even though modern HPV assays have high sensitivity, the sampling procedure of male specimens (for example, type of swab or anatomic collection site in male genitalia) is not as established as that for female sampling and may influence the detection of cellular or viral DNA. However, the missing mechanism should not depend on presence or absence of HPV [33], and techniques for dealing with data with missing at random(MAR), such as multiple imputation method [34] can be applied.

We illustrated the models using the HPV dataset from the Rakai circumcision trial to evaluate the as-treated efficacy of MC and showed that the current efficacy of MC for prevention of incident HR-HPV infections, which utilized all the data, was greater than the originally reported efficacy which used the individual as the unit of analysis. To study data of all tested types in the trial and compare type-specific incidence and persistence warrant more comprehensive analyses on this dataset in the near future. The presented modeling framework also provides an analytical method to explore cross-genotype protection in future HPV vaccine studies that target a wider range of HPV genotypes.

Acknowledgments

We thank the male circumcision trial participants and the Rakai field work teams, without whom we would not have the example dataset. The Rakai trial was supported by a grant (U01 AI11171-01-02) from the National Institutes of Allergy and Infectious Disease (NIAID), Division of AIDS, National Institutes of Health (NIH), and in part by the Division of Intramural Research, NIAID, NIH.

References

1. Morse, S.; Ballard, R.; Holmes, K.; AA, M. Atlas of Sexually Transmitted Diseases and AIDS. Elsevier Science; 2003.
2. Centers for disease control and prevention: Hpv vaccine monitoring. 2007. <http://www.cdc.gov/STD/HPV/monitoring-rpt.htm>
3. Murilo R, Molano M, Martinez G, Mejia J, Gamboa O. Age-specific prevalence, incidence, and duration of human papillomavirus infections in a cohort of 290 us men. Journal of Infectious Diseases. 2008; 98(6):827–835.
4. Giuliano A, Lu B, Nielson C, Flores R, Papenfuss M, Lee J, Abrahamsen M, Harris R. Age-specific prevalence, incidence, and duration of human papillomavirus infections in a cohort of 290 us men. Journal of Infectious Diseases. 2008; 198(6):827–835. [PubMed: 18657037]
5. Hernandez B, Shvetsov Y, Goodman M, Wilkens L, Thompson P, Zhu X, Ning L. Reduced clearance of penile human papillomavirus infection in uncircumcised men. Journal of Infectious Diseases. 2010; 201(9):1340–1343. [PubMed: 20350160]
6. Fukuchi E, Sawaya G, Chirenje M, Magure T, Tuveson J, Ma Y, Shiboski S, Da Costa M, Palefsky J, Moscicki A, et al. Cervical human papillomavirus incidence and persistence in a cohort of hiv-negative women in zimbabwe. Sexually Transmitted Diseases. 2009; 36(5):305–311. [PubMed: 19295468]
7. Banura C, Sandin S, van Doorn L, Quint W, Kleter B, Wabwire-Mangen F, Mbidde E, Weiderpass E. Type-specific incidence, clearance and predictors of cervical human papillomavirus infections (hpv) among young women: a prospective study in uganda. Infectious Agents and Cancer. 2010; 5:7. [PubMed: 20380709]

8. Nielsen A, Iftner T, Munk C, Kjaer S. Acquisition of high-risk human papillomavirus infection in a population-based cohort of danish women. *Sexually Transmitted Diseases*. 2009; 36(10):609–615. [PubMed: 19955872]
9. Tam L, Chan P, Ho S, Yu M, Yim S, Cheung T, Wong M, Li E. Natural history of cervical papilloma virus infection in systemic lupus erythematosus - a prospective cohort study. *Journal of Rheumatology*. 2009 Dec 23. [Epub ahead of print].
10. Lu B, Wu Y, Nielson C, Flores R, Abrahamsen M, Papenfuss M, Harris R, Giuliano A. Factors associated with acquisition and clearance of human papillomavirus infection in a cohort of us men: a prospective study. *Journal of Infectious Diseases*. 2009; 199(3):362–371. [PubMed: 19133808]
11. Gray R, Serwadda D, Kong X, Makumbi F, Kigozi G, Gravitt P, Watya S, Nalugoda F, Ssempijja V, Tobian A, et al. Circumcision of hiv-infected men: Effects on high risk human papillomavirus infections in a randomized trial in rakai, uganda. *Journal of Infectious Diseases*. 2010; 201(10):1455–1462. [PubMed: 20370483]
12. Serwadda D, Wawer M, Makumbi F, Kong X, Kigozi G, Gravitt P, Watya S, Nalugoda F, Ssempijja V, Tobian A, et al. Circumcision of hiv-infected men: Effects on high risk human papillomavirus infections in a randomized trial in rakai, uganda. *Journal of Infectious Diseases*. 2010; 201(10):1463–1469. [PubMed: 20370481]
13. Paramsothy P, Jamieson D, Heilig C, Schuman P, Klein R, Shah K, Rompalo A, Cu-Uvin S, Duerr A. The effect of highly active antiretroviral therapy on human papillomavirus clearance and cervical cytology. *Obstetrics and Gynecology*. 2009; 113(1):26–31. [PubMed: 19104356]
14. Rowhani-Rahbar A, Hawes S, Sow P, Toure P, Feng Q, Dem A, Dembele B, Critchlow C, N'Doye I, Kiviat N. The impact of hiv status and type on the clearance of human papillomavirus infection among senegalese women. *Journal of Infectious Diseases*. 2007; 196(6):887–894. [PubMed: 17703420]
15. Koshiol J, Schroeder J, Jamieson D, Marshall S, Duerr A, Heilig C, Shah K, Klein R, Cu-Uvin S, Schuman P, et al. Smoking and time to clearance of human papillomavirus infection in hiv-seropositive and hiv-seronegative women. *American Journal of Epidemiology*. 2006; 164(2):176–183. [PubMed: 16775041]
16. Xue X, Gange S, Zhong Y, Burk R, Minkoff H, Massad L, Watts D, Kuniholm M, Anastos K, Levine A, et al. Marginal and mixed-effects models in the analysis of human papillomavirus natural history data. *Cancer Epidemiol Biomarkers Prev*. 2010; 19(1):159–169. [PubMed: 20056635]
17. Wei L, Lin D, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *Journal of the American Association*. 1989; 84:1065–1073.
18. Kong X, Gray R, Archer K, Moulton L, Wang M. Parametric frailty models for clustered data with arbitrary censoring: Application to effect of male circumcision on hpv clearance. *BMC Medical Research Methodology*. 2010; 10:40. [PubMed: 20459614]
19. Agresti, A. *Categorical data analysis*. Wiley-Interscience; 2002.
20. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
21. Carey V, Zeger S, Diggle P. Modeling multivariate binary data with alternating logistic regressions. *Biometrika*. 1993; 80:517–526.
22. Breslow N, Clayton D. Approximate inference in generalized linear mixed models. *Journal of the American Association*. 1993; 88:9–25.
23. Bonney G. Logistic regression for dependent binary observations. *Biometrics*. 1987; 43:951–973. [PubMed: 3427178]
24. FitzGerald E, Knuiman M. Interpretation of regressive logistic regression coefficients in analyses of familial data. *Biometrics*. 1998; 54:909–920.
25. Gray R, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, Kiwanuka N, Moulton L, Chaudhary M, Chen M, et al. Male circumcision for hiv prevention in men in rakai, uganda: a randomised trial. *Lancet*. 2007; 369:657–666. [PubMed: 17321311]
26. Burchell A, Winer R, Sanjose S, Franco E. Chapter 6: Epidemiology and transmission dynamics of genital hpv infection. *Vaccine*. 2006; S3:S3/52–S3/61. [PubMed: 16950018]

27. Wedderburn R. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*. 1974; 61:439–447.
28. McCullagh P. Quasi-likelihood functions. *Annals of Statistics*. 1983; 11:59–67.
29. Zeger S, Qaqish B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics*. 1988; 44:1019–1031. [PubMed: 3148334]
30. McCullagh, P.; Nelder, J. Generalized linear models. Chapman and Hall/CRC; 1989.
31. Carey V, Zeger S,PD. Modeling multivariate binary data with alternation logistic regressions. *Biometrika*. 1993; 80:517–526.
32. Liang K, Zeger S. A class of logistic regression models for multivariate binary time series. *Journal of the American Statistical Association*. 1989; 84:447–451.
33. Bleeker M, Hogewoning C, Berkhof J, Voorhorst F, Hesselink A, Diemen van P, den van BA, Snijders P, Meijer C. Concordance of specific human papillomavirus types in sex partners is more prevalent than would be expected by chance and is associated with increased viral loads. *Clinical Infectious Diseases*. 2005; 41(5):612–620. [PubMed: 16080082]
34. Rubin, D. Multiple imputation for nonresponse in surveys. Wiley; 2004.

Figure 1. Testing results observed on one individual: an example illustrating the data in HPV empidemiology studies. ($T_0, T_1, \dots, T_3$:baseline and follow up visits).

Figure 2.
Power and estimation accuracy vs. number of simulated individuals

Table 1

As-treated efficacy of male circumcision for HR-HPV incidence and persistence

Circumcised	Incidence risk ratio (95%CI)	Persistence risk ratio (95% CI)
Yes vs. NO	0.55 (0.42, 0.73)	0.67 (0.49, 0.92)

Table 2

HR-HPV Incidence and persistence odds ratio estimates of demographical, hygiene, and behavioral factors

	Incidence odds ratio (95%CI)	Persistence odds ratio (95% CI)
Age (Ref:<=30)		
31–40	0.67 (0.50, 0.89) *	1.00 (0.75, 1.35)
>40	0.42 (0.25, 0.71) *	0.96 (0.48, 1.89)
Education (ref: No)		
Primary	1.90 (1.10, 3.30) *	0.68 (0.42, 1.10)
Secondary or higher	1.91 (1.04, 3.50) *	0.69 (0.39, 1.24)
Currently married (Ref:No)		
Yes	0.87 (0.44, 1.74)	1.03 (0.64, 1.64)
Non-marital partnership (Ref:No)		
Yes	1.56 (1.14, 2.14) *	0.85 (0.54, 1.32)
No. of sexes past 30 days		
Every 5 more	1.54 (1.23, 1.93) *	0.79 (0.48, 1.30)
No. of sex partners (Ref:single)		
Multiple	1.70 (1.29, 2.24) *	0.74 (0.51, 1.10)
Freq of washing genital (ref: <daily)		
Daily	0.97 (0.66, 1.43)	0.76 (0.44, 1.30)
More often	0.86 (0.56, 1.33)	0.93 (0.51, 1.69)
Condom use (ref:No)		
Always	1.17 (0.38, 3.56)	1.00 (0.28, 3.61)
Sometimes	1.43 (1.08, 1.90) *	0.77 (0.52, 1.64)
Alcohol use before sex (ref:no)		
Yes	1.10 (0.50, 2.41)	1.02 (0.43, 2.41)

* significant at level 0.05.