Published in final edited form as:
Stat Med. 2011 July 30; 30(17): 2160–2170. doi:10.1002/sim.4257.

Discrete-time semi-Markov modeling of human papillomavirus persistence

C. E. Mitchell[a], **M. G. Hudgens**[a,*], **C. C. King**[b], **S. Cu-Uvin**[c], **Y. Lo**[d], **A. Rompalo**[e], **J. Sobel**[f], and **J. S. Smith**[g]

[a] The University of North Carolina, Department of Biostatistics, Chapel Hill, NC

[b] Centers for Disease Control and Prevention, Atlanta, GA

[c] Brown Medical School, Providence, RI

[d] Montefiore Medical Center and Albert Einstein College of Medicine, Bronx, NY

[e] Johns Hopkins University School of Medicine, Baltimore, MD

[f] Wayne State University School of Medicine, Detroit, MI

[g] The University of North Carolina, Department of Epidemiology, Chapel Hill, NC

Abstract

Multi-state modeling is often employed to describe the progression of a disease process. In epidemiological studies of certain diseases, the disease state is typically only observed at periodic clinical visits, producing incomplete longitudinal data. In this paper we consider fitting semi-Markov models to estimate the persistence of human papillomavirus (HPV) type-specific infection in studies where the status of HPV type(s) is assessed periodically. Simulation study results are presented indicating the semi-Markov estimator is more accurate than an estimator currently used in the HPV literature. The methods are illustrated using data from the HIV Epidemiology Research Study (HERS).

Keywords

panel data; stochastic process

1. Introduction

1.1. Defining and estimating HPV persistence

Persistent HPV infection is considered to drive the progression of cervical neoplasia to cervical cancer [1, 2], the second most frequently occurring cancer in women worldwide. HPV persistence is associated with high-grade cervical intraepithelial neoplasia (CIN 2/3), or pre-cancerous lesions of grades 2 or 3, which may develop into invasive cervical cancer (ICC) [1]. Thus, HPV persistence is important as a clinical marker and endpoint in clinical trials [1] and in screening to identify women who are at highest risk of high grade pre-cancerous lesions and cervical cancer [1, 3]. Characterizing the persistence of HPV infections is also important in studying the natural disease history of cervical cancer.

* Correspondence to: 135 Dauer Drive, CB # 7420, Chapel Hill, NC 27599. mhudgens@bios.unc.edu.

Despite its importance, there is no consensus regarding what exactly constitutes a "persistent" infection [1, 4]. Conceptually HPV persistence is defined as the length of time during which an individual is infected with an HPV infection, i.e., the duration the infection persists. Because a woman can become infected with one or more types of HPV, persistence is often defined in terms of the duration of a type specific infection. Unfortunately, it is impossible to observe the duration of a type specific infection. Clinical trials and natural history studies of HPV typically produce incomplete data where type-specific HPV infection is observed intermittently at a sequence of discrete time points (study visits), resulting in unobservable exact times of transition between infection states (i.e., "panel data") [5, 6]. Even if infection status could be monitored continuously, HPV tests are typically based on viral load being above or below a detection limit. Test results below a limit of detection do not distinguish between the infection being cleared versus remaining latent in some reservoir in the body. As a result of these issues, investigators typically resort to an operational definition of persistence as a surrogate for the true unobservable underlying continuous infection process. The aim of this paper is the development of a method for estimating the distribution of type-specific persistence from longitudinal HPV test results based on whatever particular operational definition is adopted by investigators.

This paper is motivated by HERS, a longitudinal study of 1310 women in the U.S. from 1993 - 2000 who either had HIV without an AIDS-defining condition (1987 Centers for Disease Control and Prevention case definition) or were at risk of HIV infection due to injection drug use or high-risk sexual behavior. There were a maximum of 15 study visits per participant, each occurring at fixed scheduled visits approximately six months apart. Each visit included a gynecological exam, cervicovaginal lavage to collect samples for detecting HPV DNA, and Papanicolaou test screening. In an analysis of longitudinal HPV data from the first 10 visits of HERS, Koshiol et al. [7] adopted an operational definition of HPV persistence defined as the time between the date of the first type-specific positive visit and the date of the first of two consecutive visits negative for the same HPV type. The requirement of two consecutive negative tests allows for possible misclassification of test results. In particular, a single intermittent type-specific HPV-negative test is not generally believed to necessarily be indicative of a woman clearing infection and then becoming reinfected with the same HPV type [4]. Rather a single intermittent type-specific negative test may represent a false negative result, perhaps due to transient suppression of viral load below the level of detection. On the other hand, two consecutive negative tests results for the same HPV type suggest the infection may have cleared.

Estimating HPV persistence is challenging for several reasons. A method for estimating HPV persistence needs to be sufficiently flexible to accommodate various operational definitions of persistence. A persistence estimator must also accommodate the incomplete data typically produced in clinical trials and natural history studies of HPV. In these settings, HPV infection may be missing at certain time points; for example, a subject may skip a clinic visit during the study for unknown reasons, the specimen drawn may be of insufficient quality to perform a valid lab test, or the lab test result may be inconclusive. Even in the absence of missed study visits or drop out, some type-specific HPV infections may still be present at the conclusion of the study, such that some accommodation for right censoring is necessary. If the operational definition of persistence allows for an HPV positive individual to clear infection and subsequently acquire a new HPV infection, a persistence estimator would also need to accommodate the possibility of repeated infections within an individual over time.

Current analytical approaches typically employed in estimating time of persistence from panel data entail using standard survival analysis methods, which we call "empirical estimators" (EEs). In the absence of right censoring, EEs reduce to observed proportions.

For example, if there are no repeated infections, the probability of a particular HPV type persisting for at least time *t* is estimated by the proportion of study individuals infected with that type where the infection lasted *t* or longer. In the presence of right censoring, these simple estimators are extended using the Kaplan-Meier method. To deal with intermittent missing HPV results, EEs often explicitly or implicitly assume that an individual is HPV positive when a study visit is missed following a visit where an HPV positive result occurred [7, 8, 9, 10, 11, 12]. EEs sometimes exclude individuals with consecutively missing HPV results [7, 9]. Intuitively, such assumptions and exclusions will lead to bias or inefficiency when estimating the duration of infection. Indeed, a simulation study is described below confirming this intuition.

As an alternative to EEs, in this paper we consider fitting semi-Markov models to estimate type specific HPV persistence from longitudinal HPV data. Markov models are often used in modeling multi-state disease processes. However, the Markov assumption may not be appropriate in modeling HPV since the probability of transitioning between states may depend on the elapsed time in the current state. For example, the likelihood that an HPV type-specific infection will not clear is known to increase with the amount of infection time [10]. When future transitions depend upon the time spent in the current state, the stochastic process is classified as semi-Markov. In the HPV setting, semi-Markov models allow for the possibility that the probability of clearing an HPV type-specific infection may depend on how long an individual has been HPV positive.

Recently, Kang and Lagakos [13] developed methods for fitting continuous-time semi-Markov multi-state models to HPV panel data. Their methods were illustrated with a model of the natural history of oncogenic genital HPV infection in women using data from the placebo arm of an HPV vaccine trial. While Kang and Lagakos avoid the usual Markov assumption employed in multi-state modeling, their proposed methods require some strong assumptions which arise from modeling time as continuous. For instance, their methods require assuming specific “guarantee time” for transitions from particular states, i.e., a-priori specification of minimum times an individual must remain in a state (such as HPV infection) before transitioning to other states. Kang and Lagakos also make parametric assumptions about the transition time distributions.

In this paper we consider discrete-time semi-Markov models for estimating type specific HPV persistence. Discrete-time models have advantages over continuous-time models, such as not requiring the specification of guarantee times or parametric distributional assumptions. Discrete time semi-Markov models have been applied in the HPV setting previously, using either Bayesian or random effects modeling [14, 15]. Here we consider fitting discrete-time semi-Markov models using nonparametric frequentist methods that do not require specification of prior distributions or parametric assumptions as in Bayesian and random effects modeling.

1.2. Outline

The outline of the remaining sections is as follows. In Section 2 an EE of type-specific HPV persistence is illustrated using panel data. In Section 3 a maximum-likelihood estimator (MLE) of type-specific HPV persistence is proposed using a semi-Markov two-state discrete-time model. Section 4 describes a simulation study comparing the MLE and EE in settings similar to HERS. Section 5 extends the two-state model from Section 3 to a more general three-state model. In Section 6 the different estimators are applied to data from HERS. Section 7 concludes with a discussion.

2. Empirical estimator

In this section, we present an illustrative example of an EE motivated by Koshiol et al. [7]. Suppose we observe five subjects in one of two observable states (0=HPV type negative or 1=HPV type positive) at seven time points $T_0,\ldots,T_6$. The following could occur:

Subject	T_0	T_1	T_2	T_3	T_4	T_5	T_6	$\tilde{\tau}_i$
1	0	0	0	0	1	0	0	1
2	0	1	0	1	*	0	0	4
3	0	1	*	1	1	0	0	4
4	0	1	0	0	0	0	0	1
5	0	0	0	1	*	0	0	2

where * represents a missing response. In the last column of the table above $\tilde{\tau}_i$ denotes an estimate of the duration of HPV type infection for individual i. Following Koshiol et al. [7], the duration estimates assume intermittent negative responses (i.e., a single zero preceded and followed by a one) and single missing responses following a positive response are actually positive responses. The EE of the probability HPV type infection persists at least t time points equals $n^{-1}\sum_{i=1}^{n} I[\tilde{\tau}_i > t]$, where $I[\cdot]$ is the usual indicator function which equals 1 if $\cdot$ is true and 0 otherwise. The estimated times $\tilde{\tau}_i$ rely on a key assumption: HPV type positive individuals with a single missing visit would have tested HPV type positive if the visit had not been missed. Intuitively one might expect such an estimator to be biased. Indeed simulations studies in Section 4 below provide empirical evidence showing the EE is biased. In the Appendix a proof is given showing the EE is not in general a consistent estimator of the duration of infection, even if there are no missed visits.

3. Semi-Markov model

As an alternative to the empirical estimator, in this section we consider maximum-likelihood estimation of HPV type-specific persistence using a semi-Markov two-state discrete-time model for individuals starting in the same infection-free state.

3.1. Model

Let $X(\cdot) = \{X_\tau : \tau \in \{0, 1, 2,\ldots\}\}$ denote a discrete-time stochastic process with state space $S = \{0, 1\}$ where $X_\tau = 0$ denotes HPV type negative and $X_\tau = 1$ denotes HPV type positive at time τ. Assume $X_0 = 0$, i.e., all individuals are HPV type negative at time 0. Let $Y_i \in \{0, 1\}$ denote the i-th state visited by the stochastic process, and let $T_i \in \{1, 2,\ldots\}$ denote the i-th sojourn time (i.e., the amount of time that an individual stays in Y_{i-1} before transitioning to Y_i). Thus, $X(\cdot)$ is equivalent to $\{Y_1, T_1, Y_2,\ldots,Y_i, T_i,\ldots\}$. Assume $X(\cdot)$ is a time-homogeneous semi-Markov process [17, 18] such that for $j \neq k$ and $i = 1, 2,\ldots$

$$P\{Y_{i+1}=k, T_{i+1}=t|Y_1,\ldots,Y_i=j;T_1,\ldots,T_i\} = P\{Y_{i+1}=k, T_{i+1}=t|Y_i=j\}, \quad (1)$$

i.e., the probability of transitioning to state k after sojourn time t in state j is independent of the history of the process. Let $p_{jk}(t) = P\{Y_{i+1} = k, T_{i+1} = t|Y_i = j\}$ for $j \neq k$ and let $p_{jj}(t)$ denote the conditional probability of remaining in state j after sojourn time t, i.e., $p_{jj}(t) = 1 - \Sigma_{j\neq k}p_{jk}(t)$.

A special case of a semi-Markov process is when future transitions depend only upon the current state, independent of time. The stochastic process is classified as Markov if

$$P\{X_{i+1}=k|X_i=j, X_{i-1}, \ldots, X_0\} = P\{X_{i+1}=k|X_i=j\}. \quad (2)$$

Let $p_{jk} = P\{X_{i+1} = k|X_i = j\}$. The Markov property (2) implies $p_{jk}(t) = p_{jk}(1 - p_{jk})^{t-1}$, i.e., the sojourn time follows a geometric distribution.

Under the Markov assumption, the stochastic process is governed by fewer parameters (p_{01} and p_{10}) than under the semi-Markov assumption. Thus, Markov models are easier to fit, but may not be flexible enough to adequately model complex processes. For example, there is evidence that the probability of clearing an HPV type infection depends on how long an individual has been infected [10]. On the other hand, there may be no reason to believe that the probability of acquiring an HPV type infection depends on how long the individual has been uninfected. Following Kang and Lagakos [13], for now we assume that the stochastic process leaving state 1 (HPV type positive) is semi-Markov and leaving state 0 (HPV type negative) is Markov [13]. That is, we assume (1) holds for $j = 1$, $k = 0$ and that (2) holds for $j = 0$, $k = 1$. Models relaxing this assumption are considered in Section 5.

Typically in HPV studies an individual's disease process is only observed until some time point at which follow-up ends. Let n_t represent the total number of possible observed time points after study entry. Under the assumption above, the observable process $X(\cdot)$ is characterized by the $(n_t + 1)$-dimensional vector $p = (p_{01}, p_{10}(1), p_{10}(2),\ldots,p_{10}(n_t - 1), p_{1+}(n_t))$ where in general

$$p_{j+}(t) = 1 - \sum_{i=1}^{t-1} p_{j(1-j)}(i)$$

for $j \in \{0, 1\}$. Let the random variable M (with realization m) denote the total number of states visited by time n_t such that Y_M is the state occupied at n_t. Let t_{M+} denote the sojourn time in Y_M at n_t, i.e., t_{M+} is the amount of time an individual has occupied Y_M at the end of the study. Let $x = (x_0, x_1,\ldots,x_{n_t})$ denote the path up to time n_t and let $\pi_{\mathbf{x}}(\boldsymbol{p}) = P\{(X_0, X_1, \ldots,X_{n_t}) = \boldsymbol{x}\}$. Then

$$\pi_x(p) = p_{y_0y_1}(t_1)\, p_{y_1y_2}(t_2) \ldots p_{y_{m-2}y_{m-1}}(t_{m-1})\, p_{y_m+}(t_{m+}) = \left\{\prod_{i=0}^{m-2} p_{y_iy_{i+1}}(t_{i+1})\right\} p_{y_m+}(t_{m+}). \quad (3)$$

3.2. Estimands

Below we show that HPV type-specific persistence at any particular time point may be written as a function of $\boldsymbol{p}$. Here the operational definition of persistence discussed in Section 1 is adopted. However, other definitions of persistence could also be used provided the estimand can be written as a function of the transition probabilities.

Let $\phi_j(\boldsymbol{p})$ denote the probability an individual is HPV type positive for j units of time followed by two HPV type negative tests allowing for single intermittent negative results. For example, if $j = 1$ then the probability HPV persists 1 unit of time is

$$\phi_1(p) = P\{X_{(0:3)} = (0100)|X_0 = 0, X_1 = 1\} = p_{10}(1)(1 - p_{01}),$$

where $X_{(0:n)} \equiv (X_0, X_1, \ldots, X_n)$. Similarly, for $j = 2$ and $j = 3$, the probabilities HPV persists 2 or 3 units of time equal

$$\phi_2(\mathrm{p}) = P\{X_{(0:4)} = (01100) | X_0 = 0, X_1 = 1\} = p_{10}(2)(1 - p_{01})$$

and

$$\begin{aligned} \phi_3(\mathrm{p}) &= P\{X_{(0:5)} = (011100) | X_0 = 0, X_1 = 1\} \\ &+ P\{X_{(0:5)} = (010100) | X_0 = 0, X_1 = 1\} \\ &= p_{10}(3)(1 - p_{01}) + p_{10}(1)^2 p_{01}(1 - p_{01}) \end{aligned}$$

respectively. More generally, define for $\boldsymbol{x}_{(0:j+2)} \in [0, 1]^{j+3}$, a vector of length $j + 3$ zeroes and ones,

$$\eta_{\mathrm{x}_{(0:j+2)}} = \begin{cases} 1 & \text{if } x_1 = x_j = 1; x_i + x_{i+1} > 0 \forall i = 2, \ldots, j-2; x_{j+1} = x_{j+2} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

That is, $\eta \boldsymbol{x}_{(0:j+2)}$ is an indicator function that equals 1 if an individual becomes HPV type positive at time 1 and does not have two consecutive HPV type negative results until time $j + 2$; otherwise, $\eta \boldsymbol{x}_{(0:j+2)}$ equals zero. Then for $j = 1, 2, \ldots$

$$\phi_j(\mathrm{p}) = \sum_{\mathrm{x}_{(0:j+2)} \in [0,1]^{j+3}} \eta \mathrm{x}_{(0:j+2)} P\left\{X_{(0:j+2)} = x_{(0:j+2)} | X_0 = 0, X_1 = 1\right\}, \tag{5}$$

and the probability type-specific HPV persists at least j units of time is

$$1 - \sum_{i=1}^{j-1} \phi_i(\mathrm{p}). \tag{6}$$

3.3. Inference

Maximum likelihood methods can be used to draw inference regarding the estimands of interest, e.g., (6). One challenge in the analysis of longitudinal HPV studies is missing data. That is, $\boldsymbol{x}$ may not be completely observable for some individuals. Assuming the missing data mechanism is missing at random (MAR) [19] (i.e., given the observed data, the missingness mechanism does not depend on the unobserved data), the likelihood contribution for an individual is obtained by summing over all possible trajectories consistent with that individual's observed data. For individual i, let $\alpha_{i\boldsymbol{x}} = 1$ if the path $\boldsymbol{x}$ is consistent with the observed data for that individual, and 0 otherwise. Revisiting the example in Section 2, for subject 1, $\alpha_{1(0000100)} = 1$ and $\alpha_{1\boldsymbol{x}} = 0$ for all $\boldsymbol{x} \neq (0000100)$; for subject 5, $\alpha_{5(0001100)} = \alpha_{5(0001000)} = 1$ and $\alpha_{5\boldsymbol{x}} = 0$ otherwise.

Under MAR the likelihood can be written as

$$L(\mathrm{p})=\prod_{i}\sum_{\mathrm{x}\in[0,1]^{n_t+1}}\alpha_{i\mathrm{x}}\pi_{\mathrm{x}}(\mathrm{p}) \tag{7}$$

where the product is over all individuals and the sum is over all possible paths of length n_t + 1. Let Ω denote the set of $\boldsymbol{p}$ satisfying the constraints

$$\Omega=\left\{\mathrm{p}:0\le p_{01}\le 1, 0\le p_{10}(t)\le 1 \text{ for } t\in\{1,\ \ldots, n_t-1\}, 0\le p_{1+}(n_t)\le 1, \sum_{t=1}^{n_t-1}p_{10}(t)+p_{1+}(n_t)=1\right\} \tag{8}$$

The MLE of $\boldsymbol{p}$ is obtained by maximizing the log-likelihood, log $L(\boldsymbol{p})$, over the parameter space Ω. The MLE of $\boldsymbol{p}$ in turn gives rise to the MLE of functions of $\boldsymbol{p}$ (e.g., $\phi_j(\boldsymbol{p})$) due to the invariance property of maximum likelihood [20].

Standard numerical methods for maximizing functions with linear equality and inequality constraints can be used to maximize log $L(\boldsymbol{p})$ over Ω. Many of these optimizers are readily available in software. For example, the SAS Version 9.2 IML procedure (SAS, Inc., Cary, North Carolina) offers a number of optimization subroutines for maximizing continuous non-linear functions subject to linear equality and inequality constraints. In the results below, the NLPQN( ) function, a (dual) quasi-Newton algorithm, was used to maximize log $L(\boldsymbol{p})$ over Ω.

Confidence intervals (CIs) for inference regarding the transition probabilities (as well as functions thereof such as (6)) can be obtained using profile likelihood [21]. A likelihood ratio test can be used to assess whether a simpler model assuming both states are Markov adequately fits the data. In particular, the likelihood can be maximized under the full model described above and under the null model $H_0: p_{10}(j) = \{1 - p_{10}(1)\}^{j-1} p_{10}(1)$ for $j = 1,\ldots, n_t - 1$. Under the H_0, the corresponding likelihood ratio test statistic will have approximately a χ^2 distribution with $n_t - 2$ degrees of freedom.

4. Simulation study

A simulation study was conducted to assess the bias of the EE and semi-Markov MLE of persistence described above. For the simulation study, time of persistence was defined as the time from first HPV type-positive result until the time of the first of two consecutive HPV type-negative results. Data sets, each with a sample size of 500 individuals, were randomly generated to be similar to the HERS data analyzed by Koshiol et al. [7]. The number of possible study visits was 10 (including study entry, such that $n_t = 9$) with two visits per year. The stochastic process leaving state 0 (HPV negative) was Markov and leaving state 1 (HPV positive) was semi-Markov. A complete set of study visits ranging from visit 0 (study entry) to visit 9 was first created for each subject. Next, to construct a pattern of missing responses attributable to drop out, a woman was right-censored with probability 0.05 at visits 2-8 and probability 0.10 at visit 9. Intermittent missing response probabilities were 0.12, 0.09, 0.11, 0.10, 0.11, 0.11, 0.08, and 0.09 for study visits 2-9, respectively.

Simulations were done under three different transition probability scenarios based on fitting the two-state model from Section 3 to the HERS data for HPV types 16 (scenario I), 53 (scenario II), and 18 (scenario III). For scenario I, the transition probabilities were $\boldsymbol{p}$ =(0.016, 0.653, 0.151, 0.085, 0.0001, 0.0001, 0.028, 0.0001, 0.0001, 0.083). Based on equation (6), under scenario I, the probabilities HPV persists at least $j = 1,\ldots, 7$ units of time

(corresponding to 0.5 to 3.5 years) were 0.36, 0.21, 0.12, 0.12, 0.11, 0.09, and 0.08. For scenario II, $\boldsymbol{p}$ =(0.038, 0.526, 0.225, 0.094, 0.059, 0.0001, 0.014, 0.0001, 0.042, 0.042). For scenario III, $\boldsymbol{p}$ =(0.015, 0.570, 0.247, 0.076, 0.0001, 0.024, 0.0001, 0.0001, 0.0001, 0.081).

For each scenario, 1000 data sets were generated. For each simulated data set, the MLE and EE of the probability of HPV infection persisting at least t years ($t \in \{0.5, 1.0, \ldots, 3.5\}$) were evaluated. The MLE was computed based on equation (6). Here and in the sequel the EE was evaluated using the Kaplan-Meier estimator as described in Koshiol et al. [7]. In particular, to compute the EE the duration of individual type-specific infections was calculated as the time between the first HPV positive visit and the first of two consecutive HPV negative visits. The previous HPV result was carried forward for single intermittent missing HPV results. HPV infections followed by a single HPV negative result at an individual's last study visit were censored at the last visit. Six months was added to the duration for HPV infections that were positive at the final visit. Women with missing HPV results at two or more consecutive visits were excluded.

The average estimates over the 1000 simulations for scenario I are presented in Figure 1. As expected, the EE tended to over-estimate the probability of HPV persistence. On the other hand, the MLEs were approximately unbiased. Results from scenarios II and III were similar (not shown). For each simulated data set, 95% profile likelihood-based CIs associated with the MLEs were calculated. Empirical coverage for each scenario was calculated by the proportion of simulations where the CI overlapped the true probability HPV persists at least $j = 1, \ldots, 7$ units of time. The empirical coverage of the profile likelihood-based CIs given in Table 1 indicates approximate nominal coverage.

5. Extensions

In the semi-Markov model developed in Section 3, a woman can occupy one of two possible states: HPV type negative (state 0) or HPV type positive (state 1), where state 0 is assumed to be Markov and state 1 is allowed to be semi-Markov. Extensions of this two-state model are considered in this section. Note the two-state model makes no distinction between (i) women who have never been infected during the study and (ii) women who have been infected during the study but subsequently cleared infection. For both (i) and (ii) the probability of infection in the two-state model equals p_{01}, i.e., no distinction is made between the probability of the initial HPV infection (after time 0) and the probability of subsequent infections. To allow for such a distinction, the two-state model can be extended to a three-state model with state space $S = \{0^*, 0, 1\}$, where state 0* denotes being HPV type negative with no prior type-specific infection (since time 0), 0 denotes being HPV type negative after an HPV type infection, and as before 1 denotes HPV type positive. Assume all individuals are in state 0* at time 0, that states 0* and 0 are Markov, and that individuals may transition from states 0* to 1, from 1 to 0, and from 0 to 1. Then the likelihood development and inferential procedures for the three-state model are analogous to those in Section 3, except there is one additional parameter to estimate, namely the probability of transition from state 0* to 1, denoted by p_{0^*1}.

The two-state model can be viewed as a special case of the three-state model where $p_{01} = p_{0^*1}$. Thus the three-state model can be used to assess the fit of the two-state model using a likelihood ratio test comparing the two models. Under the null hypothesis the two-state model holds, i.e., $p_{01} = p_{0^*1}$, the likelihood ratio test statistic will have approximately a χ^2 distribution with 1 degree of freedom.

The three-state model can be generalized even further by letting state 0 be semi-Markov. Likelihood development and inference are again similar to Section 3, except the single parameter p_{01} is replaced by $n_t - 1$ parameters $p_{01}(1), p_{01}(2), \ldots, p_{0+}(n_t - 1)$. Note the

transition probabilities from state 0 to state 1 are only identifiable from the observable data for sojourn times up to $n_t - 2$ since all individuals are assumed to begin in state 0*. For the estimands defined in Section 3.2, p_{01} is replaced by $p_{01}(1)$. Based on this more general three-state model a likelihood ratio test can be employed to assess whether the simpler three-state model assuming state 0 is Markov adequately fits the data, with the test statistic having approximately a χ^2 distribution with $n_t - 3$ degrees of freedom (assuming $n_t > 3$). Further generalization of the three state model allowing for state 0* to be semi-Markov is more difficult as the duration a woman has occupied state 0* prior to the start of a study will in general not be known.

6. HIV Epidemiology Research Study (HERS)

In this section we analyze data from the HERS cohort described in Section 1.1. Women were included in this analysis if they were HIV seropositive at baseline and had two or more study visits, HPV DNA results at study entry, a cervix, and no cervical treatment in the past 6 months prior to enrollment. Only data from the first 10 visits were analyzed. At each visit in HERS women were tested for 26 possible HPV types. For this analysis, HPV type 53, 16, and 18 infections were analyzed separately. HPV type 53 was analyzed because it was the most common individual type among HIV positive women enrolled in the study [7]; HPV types 16 and 18 are the two most common cancer-associated HPV types worldwide [22]. Each type-specific analysis only included women who were HPV type-specific negative at study enrollment. For instance, the HPV type 16 analysis included only the 524 women who were not infected with HPV 16 at study enrollment. Similarly, there were 516 (500) women who were not infected with HPV type 18 (53) at enrollment.

Type-specific persistence was estimated from the HERS data using the EE and the semi-Markov models. The EE was computed as described above. In order to fit the discrete time semi-Markov models to the HERS data, visit times were rounded to the nearest six month scheduled time point; visit times were not rounded for calculating the EE. The two-state model from Section 3 and the three-state models from Section 5 were fit separately for each type. For type 16 the three-state model assuming states 0* and 0 are Markov and state 1 is semi-Markov provided a better fit than the two-state model (likelihood ratio test p-value < 0.001). Similar results were obtained for type 53 (p-value < 0.001) and type 18 (p-value < 0.001). For type 16 there was no improvement in fit by allowing state 0 to be semi-Markov (p-value = 0.23). Similar results were obtained for type 53 (p-value = 0.12). For type 18 the more general three-state model provided a slightly better fit (p-value = 0.01), suggesting the probability of reinfection with type 18 may be time dependent. Estimates of type-specific persistence for the best fitting models are given in Table 2 for all types and in Figure 2 for type 16. As expected, estimates using the EE are higher than from the semi-Markov models.

7. Conclusion

This research was motivated by an analysis of the HERS cohort to estimate HPV type-specific persistence. Our results suggest EEs may result in overestimation of persistence of HPV type infections. In general, EEs are not consistent, and simulation studies demonstrate substantial bias of EEs in finite samples. Alternatively, using discrete-time semi-Markov models, we consider a maximum likelihood-based estimator of HPV type-specific persistence. If the model assumptions are correct, the resulting MLEs will in general be consistent estimators of persistence; simulation studies indicate the MLEs are approximately unbiased. Comparison of the EE and MLE applied to the HERS data indicates the bias of the EE can be large in practice.

There are several appealing aspects of the MLE of HPV type-specific persistence. First, this estimator requires no parametric distributional assumptions or a priori specification of guarantee times. Second, the underlying semi-Markov model allows the probability of clearing an HPV type infection to possibly depend on the time infected with that type. Third, the MLE gives valid large sample inference when the HPV infection model is correctly specified and the missing data mechanism is MAR. Simulation results suggest the MLE is approximately unbiased in finite sample settings similar to HERS. Finally, the method is flexible with respect to estimands. Namely, an estimator of any definition of persistence can be constructed provided the estimand can be written as a function of the transition probability estimates. In other words, if an investigator wishes to consider more than one persistence definition (e.g., time until first of two negative tests compared to time until first negative test), the MLEs of both types of persistence are easily computed as functions of the MLEs of the transition probabilities.

The proposed method depends on time being sufficiently discrete, which can be achieved by rounding or coarsening of observation times for studies that do not have planned visit schedules. This discretization may lead to bias or loss of precision. However, without coarsening the observation times, the number of parameters in the semi-Markov model may become prohibitively large. Thus, this method may be best suited for studies with regularly scheduled visits (e.g., clinical trials), providing a non-parametric approach to summarizing the observable discrete infection data from such studies. If, on the other hand, the goal is to extrapolate from the observable data to the underlying unobservable continuous infection process, then the methods proposed by Kang and Lagakos [13] may be more appropriate. By treating time as continuous, the Kang and Lagakos method does not require discrete visit times. Their method does however require certain strong assumptions, such as parametric distributions and guarantee times, that the discrete time model does not require.

There are many possible avenues of further methodological research related to estimating HPV persistence. For instance, both the proposed method and the Kang and Lagakos approach assume all individuals are HPV type negative at baseline. However, in longitudinal studies of HPV, a woman may already be infected at the first study visit, i.e., at study enrollment. Further research is needed to allow for such “prevalent infections.” Additional research is also needed on methods to combine data across HPV studies, where visit schedules and patient characteristics may differ between studies.

Finally we note that while motivated by HPV, the discrete time semi-Markov model could be applied to other settings where recurrent infections or re-activations occur, such as malaria, herpes, or parasitic infection [23, 24, 25], and only panel data is available on the infection/activation state.

Acknowledgments

Contract/grant sponsor: NCI R01 CA114773-05

Appendix

Here we sketch a proof that the EE is not, in general, a consistent estimator of the duration of infection. Consider the case where there are three follow-up time points (i.e., $n_t = 3$) and there is no missing data. The table below shows the eight possible observed data patterns and for each pattern the corresponding estimated duration of HPV infection $\tilde{\tau}$ used in computing the EE (as described in Sections 2 and 4).

$\boldsymbol{x} = (x_0, x_1, x_2, x_3)$	$\tilde{\tau}$

(0100)	1
(0111)	3+
(0101)	3+
(0110)	2+
(0010)	1+
(0011)	2+
(0001)	1+
(0000)	*

Here, for $j \in \{1, 2, 3\}$, we let $j+$ denote scenarios where the duration of infection is right censored in the sense that from the observed data the duration is known to be at least j units of time and possibly longer. For example, for $\boldsymbol{x} = (0111)$, the duration of infection is at least 3 time units. Because duration of infection is defined as the time between the first type-specific HPV positive visit and the first of two consecutive type-specific HPV negative visits, for $\boldsymbol{x} = (0101)$ the duration of infection is also at least 3 time units. Note in the last row that $\boldsymbol{x} = (0000)$ does not have an estimated duration of HPV infection $\tilde{\tau}$ since individuals with this observed data pattern are never infected during follow-up and thus contribute no information about the duration of infection.

As mentioned in Section 4, Koshiol et al. added one time unit to the estimated duration of infection for HPV infections that were positive at the final visit. This convention is not employed in the table above, but were this convention adopted the proof below remains unchanged.

Now consider estimating the probability an infection is of duration 1 time unit (i.e., an individual is HPV positive for one time unit followed by two consecutive negative tests). Under the semi-Markov model, this probability equals $\phi(1) = p_{10}(1)(1 - p_{01})$. Below we show that the EE of the probability an infection is of duration 1 does not in general converge in probability to $\phi(1)$. In particular, for a data set of individuals with three follow-up time points and no missing data, the Kaplan-Meier estimator of the probability of an infection having duration 1 equals

$$\tilde{\phi}(1) = \frac{\sum_i I[\mathrm{x}_i = (0100)]}{\sum_i I[\mathrm{x}_i \neq (0000)]}, \tag{9}$$

where the summation is over all individuals. The numerator of (9) is the number of individuals with an infection of duration 1 time unit and the denominator of (9) is the number of individuals with an infection of duration at least 1 time unit. By the weak law of large numbers and Slutsky's theorem,

$$\tilde{\phi}(1) \to^{p} \frac{P[\mathrm{x} = (0100)]}{1 - P[\mathrm{x} = (0000)]} = \frac{p_{01}p_{10}(1)(1 - p_{01})}{1 - (1 - p_{01})^3}. \tag{10}$$

Clearly the right side of (10) does not in general equal $\phi(1) = p_{10}(1)(1 - p_{01})$. In fact, the right side of (10) will always be less than $\phi(1)$ provided $0 < p_{01} < 1$.

To illustrate the extent of the bias that can occur by using the EE, suppose $p_{01} = 0.016$ and $p_{10}(1) = 0.653$. Then the probability an infection is of duration 1 equals $\phi(1) = 0.643$. Yet as the sample size tends to infinity, the EE $\tilde{\phi}(1)$ will converge in probability to 0.218.

References

1. Koshiol J, Lindsay L, Pimenta JM, Poole C, Jenkins D, Smith JS. Persistent human papillomavirus infection and cervical neoplasia: a systematic review and meta-analysis. Am J Epidemiol. 2008; 168:123–137. [PubMed: 18483125]
2. Schiffman MH, Bauer HM, Hoover RN, Glass AG, Cadell DM, Rush BB, Scott DR, Sherman ME, Kurman RJ, Wacholder S. Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia. J Natl Cancer Inst. 1993; 85:958–964. [PubMed: 8388478]
3. Castle PE. Invited commentary: is monitoring of human papillomavirus infection for viral persistence ready for use in cervical cancer screening? Am J Epidemiol. 2008; 168:138–144. [PubMed: 18483124]
4. Woodman CB, Collins SI, Young LS. The natural history of cervical HPV infection: unresolved issues. Nat Rev Cancer. 2007; 7:11–22. [PubMed: 17186016]
5. Gentleman RC, Lawless JF, Lindsey JC, Yan P. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. Statistics in Medicine. 1994; 13:805–821. [PubMed: 7914028]
6. Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. Journal of the American Statistical Association. 1985; 80:863–871.
7. Koshiol JE, Schroeder JC, Jamieson DJ, Marshall SW, Duerr A, Heilig CM, Shah KV, Klein RS, Cu-Uvin S, Schuman P, Celentano D, Smith JS. Time to clearance of human papillomavirus infection by type and human immunodeficiency virus serostatus. Int J Cancer. 2006; 119:1623–1629. [PubMed: 16646070]
8. Rostich A, Koshiol J, Poole C, Hudgens M, Franco E, Backes D, Pimenta J, Smith JS. Genital human papillomavirus persistence patterns among women worldwide: a literature review and meta-analysis. 2011 In preparation.
9. Koshiol J, Schroeder J, Jamieson DJ, Marshall SW, Duerr A, Heilig CM, Shah KV, Klein RS, Cu-Uvin S, Schuman P, Celentano D, Smith JS. Smoking and time to clearance of human papillomavirus infection in HIV-seropositive and HIV-seronegative Women. Am J Epidemiol. 2006; 164:176–183. [PubMed: 16775041]
10. Trottier H, Mahmud S, Prado JC, Sobrinho JS, Costa MC, Rohan TE, Villa LL, Franco EL. Type-specific duration of human papillomavirus infection: implications for human papillomavirus screening and vaccination. J Infect Dis. 2008; 197:1436–1447. [PubMed: 18419547]
11. Liaw KL, Glass AG, Manos MM, Greer CE, Scott DR, Sherman M, Burk RD, Kurman RJ, Wacholder S, Rush BB, Cadell DM, Lawler P, Tabor D, Schiffman M. Detection of human papillomavirus DNA in cytologically normal women and subsequent cervical squamous intraepithelial lesions. J Natl Cancer Inst. 1999; 91:954–960. [PubMed: 10359548]
12. Woodman CB, Collins S, Winter H, Bailey A, Ellis J, Prior P, Yates M, Rollason TP, Young LS. Natural history of cervical human papillomavirus infection in young women: a longitudinal cohort study. Lancet. 2001; 357:1831–1836. [PubMed: 11410191]
13. Kang M, Lagakos SW. Statistical methods for panel data from a semi-Markov process, with application to HPV. Biostatistics. 2007; 8:252–264. [PubMed: 16740624]
14. Plummer M, Schiffman M, Castle P, Maucort-Boulch D, Wheeler CM. for the ALTS (Atypical Squamous Cells of Undetermined Significance/Low-Grade Squamous Intraepithelial Lesions Triage Study) Group. A 2-year prospective study of human papillomavirus persistence among women with a cytological diagnosis of atypical squamous cells of undetermined significance or low-grade squamous intraepithelial lesion. J Infect Dis. 2007; 195:1582–1589. [PubMed: 17471427]
15. Maucort-Boulch D, Plummer M, Castle PE, Demuth F, Safaeian M, Wheeler CM, Schiffman M. Predictors of human papillomavirus persistence among women with equivocal or mildly abnormal cytology. Int J Cancer. 2010; 126:684–691. [PubMed: 19609952]
16. Ahdieh L, Klein RS, Burk R, Cu-Uvin S, Schuman P, Duerr A, Safaeian M, Astemborski J, Daniel R, Shah K. Prevalence, incidence, and type-specific persistence of human papillomavirus in human immunodeficiency virus (HIV)-positive and HIV-negative women. J Infect Dis. 2001; 184:682–690. [PubMed: 11517428]

17. Barbu V, Boussemart M, Limnios N. Discrete-time semi-Markov model for reliability and survival analysis. Communications in Statistics: Theory and Methods. 2004; 33(11):2833–2868.
18. Barbu V, Limnios N. Empirical estimation for discrete-time semi-Markov processes with applications in reliability. Journal of Nonparametric Statistics. 2006; 18(7-8):483–498.
19. Little, RJA.; Rubin, DB. Statistical Analysis With Missing Data. Wiley; 1987.
20. Casella, G.; Berger, RL. Statistical Inference. second. Thomson Learning, S.Pacific Grove; Calif.: 2002.
21. Murphy SA, van der Vaart AW. On profile likelihood. Journal of the American Statistical Association. 2000; 95:449–465.
22. Clifford GM, Smith JS, Plummer M, Munoz N, Franceschi S. Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. Br J Cancer. 2003; 88:63–73. [PubMed: 12556961]
23. Nagelkerke NJ, Chunge RN, Kinoti SN. Estimation of parasitic infection dynamics when detectability is imperfect. Stat Med. 1990; 9:1211–1219. [PubMed: 2247721]
24. Crespi CM, Cumberland WG, Blower S. A queueing model for chronic recurrent conditions under panel observation. Biometrics. 2005; 61:193–198. [PubMed: 15737093]
25. Singer B, Cohen JE. Estimating malaria incidence and recovery rates from panel surveys. Mathematical Biosciences. 1980; 49:273–305.

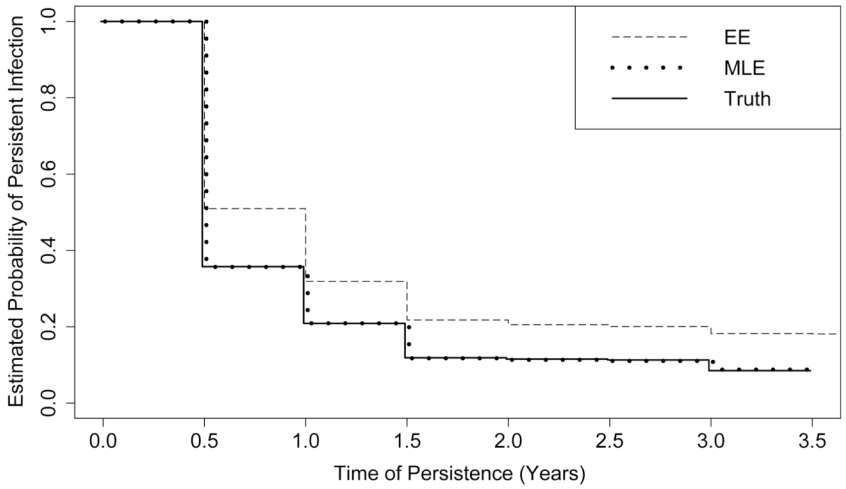

Figure 1.
Mean of empirical estimator (EE) and maximum likelihood estimator (MLE) from simulation study scenario 1 (***p*** = (0.016, 0.653, 0.151, 0.085, 0.0001, 0.0001, 0.028, 0.0001, 0.0001, 0.083))

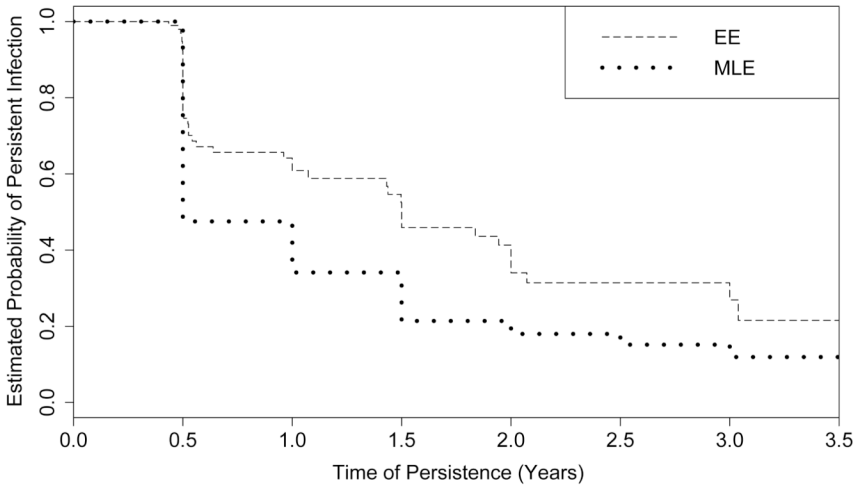

Figure 2.
Empirical estimate (EE) and maximum likelihood estimate (MLE) of HPV type-16 persistence among women in HERS cohort who were HIV positive and HPV negative at study entry.

Table 1

Empirical coverage of profile likelihood 95% confidence intervals of the probability HPV infection persists at least *t* years. Coverage is based on 1000 simulations per scenario.

	Scenario I		Scenario II		Scenario III	
t **(years)**	**Truth**	**Coverage**	**Truth**	**Coverage**	**Truth**	**Coverage**
0.5	0.36	0.95	0.49	0.95	0.44	0.95
1.0	0.21	0.95	0.28	0.94	0.20	0.95
1.5	0.12	0.96	0.18	0.95	0.12	0.95
2.0	0.12	0.96	0.11	0.95	0.11	0.95
2.5	0.11	0.96	0.11	0.95	0.09	0.95
3.0	0.09	0.96	0.09	0.94	0.08	0.94
3.5	0.08	0.96	0.09	0.95	0.08	0.94

Table 2

Estimated probabilities of type-specific HPV infection persisting at least *t* years for the empirical estimator (EE) and semi-Markov maximum likelihood estimator (MLE) based on women in HERS cohort who were HIV positive and HPV negative at study entry

	Type 16* (n=49)		Type 53* (n=102)		Type 18** (n=48)	
t (years)	EE	MLE†	EE	MLE†	EE	MLE†
0.5	0.84	0.48 [0.35, 0.61]	0.85	0.60 [0.51, 0.68]	0.92	0.48 [0.33, 0.64]
1.0	0.61	0.34 [0.23, 0.47]	0.58	0.40 [0.32, 0.48]	0.64	0.24 [0, 0.41]
1.5	0.44	0.21 [0.12, 0.34]	0.45	0.28 [0.21, 0.36]	0.48	0.13 [0, 0.28]
2.0	0.35	0.18 [0.09, 0.30]	0.43	0.19 [0.13, 0.27]	0.43	0.10 [0, 0.22]
2.5	0.30	0.15 [0, 0.28]	0.39	0.16 [0.10, 0.24]	0.25	0.07 [0, 0.19]
3.0	0.30	0.12 [0, 0.24]	0.31	0.13 [0.08, 0.21]	0.25	0.06 [0, 0.18]
3.5	0.20	0.11 [0, 0.24]	0.29	0.12 [0.07, 0.20]	0.25	0.06 [0, 0.17]

Three-state model assuming states 0 and 0 are Markov and state 1 is semi-Markov

**Three-state model assuming state 0* is Markov and states 0 and 1 are semi-Markov

†Profile likelihood 95% confidence intervals in [,]