SEMI-MARKOV MULTI-STATE MODELING OF HUMAN PAPILLOMAVIRUS

by
Cicely E. Mitchell

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics.

Chapel Hill
2011

Approved by:

Dr. Michael Hudgens, Advisor
Dr. Jianwen Cai, Member
Dr. Vidyadhar Kulkarni, Member
Dr. Devan Mehrotra, Member
Dr. Jennifer Smith, Member
Abstract

CICELY E. MITCHELL: SEMI-MARKOV MULTI-STATE MODELING OF HUMAN PAPILLOMAVIRUS.
(Under the direction of Dr. Michael Hudgens.)

In the progression of certain diseases, such as cervical neoplasia, individuals transition between a set of well-defined states over the course of time. Epidemiological natural history studies of such diseases typically produce incomplete longitudinal data, since the disease states are observed only at fixed clinic visits. Multi-state modeling is often employed in these settings to estimate various parameters of the disease process from the observed data. For example, in studying the etiology of invasive cervical cancer (ICC), multi-state models are often used to characterize the transition from human papillomavirus (HPV) acquisition to the development of cervical disease or cervical intraepithelial neoplasia (CIN) (i.e., pre-cancerous lesions as diagnosed by histology), and, ultimately, ICC. The purpose of this dissertation is to study multi-state modeling of HPV infection. We propose to develop methods for fitting such models under minimal assumptions that are more realistic than the Markov assumptions typically invoked when fitting multi-state models to data from HPV studies. The developed methods will provide a natural way to combine data from both incident and prevalent infections. Mixed effects models for infections with multiple HPV types will also be developed. The proposed methods will be compared with analytic approaches often employed in the HPV literature. The proposed methodology will be illustrated using longitudinal data from the HIV Epidemiology Research Study (HERS, United States, 1993 - 2000) and from phase IIa and III clinical trials of a HPV vaccine.
Acknowledgments

I would like to thank Dr. Hudgens for guidance and pushing me further than anyone has before, my dissertation committee for time and attention, Dr. Lloyd Edwards for advising me, Jeff Veach for inspiring me to come back to school, Ahinee and Kyna, Mom and Dad, God, and last but not least Albert for being my loving life partner and biggest fan.
# Table of Contents

Abstract ........................................................................................................ iii

List of Tables ..................................................................................................... viii

List of Figures ................................................................................................... ix

1 INTRODUCTION ............................................................................................ 1

2 BACKGROUND ................................................................................................ 4

  2.1 The Role of HPV Persistence in Cervical Cancer ....................................... 4

  2.2 Defining Persistence ................................................................................... 5

  2.3 Panel Data .................................................................................................. 6

  2.4 Existing Estimators of Persistence ............................................................. 7

  2.5 Stochastic Models ..................................................................................... 9

  2.6 Summary and Proposed Research ............................................................. 13

3 SEMI-MARKOV MODELING OF HPV PERSISTENCE .................................... 14

  3.1 Introduction ............................................................................................... 14

    3.1.1 Defining and estimating HPV persistence ........................................... 14

    3.1.2 Outline ............................................................................................... 18
List of Tables

3.1 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. ................................................. 32

3.2 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 ............................................. 33

4.1 Estimated probability HPV persists at least $t$ years for the Incidence Only (IO), Naive (NSME), and SSPLA semi-Markov maximum likelihood estimators based on the 571 women in HERS cohort who were HIV positive at study entry with at least 2 study visits .................................................. 49

4.2 Estimated bias, MSE, and RE of the probability HPV infection persists at least $t$ years from the simulation study described in section 4.5. ..................... 50

5.1 Frequency and percentage of high risk human papillomavirus types from the 571 HIV-seropositive females in the HIV Epidemiology Research Study with at least two visits .................................................. 69

5.2 Model selection for HERS analysis .................................................. 70

5.3 Frequency and percentage of human papillomavirus types from the placebo-arm of two vaccine clinical trials for women that were HPV negative for all types at baseline and had two or more visits .................................................. 71

5.4 Model selection for analysis of Protocols 005 and 012 .......................... 72

5.5 Estimated bias of the probability HPV infection persists at least $t$ years from the simulation study described in section 5.4. .......................... 73

5.6 Estimated bias from the simulation study described in section 5.4. .......................... 74
List of Figures

3.1 Mean of empirical estimator (EE) and maximum likelihood estimator (MLE) from simulation study scenario 1 ($p=(0.016, 0.653, 0.151, 0.085, 0.0001, 0.0001, 0.028, 0.0001, 0.0001, 0.083)$) .......................................................... 34

3.2 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. ......................................................... 35

4.1 Random initial observation of a two-state stochastic process .......................... 48

4.2 Model to differentiate HPV positive at study entry (state 2) from HPV positive following HPV negative (state 1) ................................................................. 51

4.3 Incidence Only (IO), naive (NSME), and Satten and Sternberg-Plummer adapted (SSPLA) semi-Markov maximum likelihood estimators (MLEs) of any HPV type persistence among women in HERS cohort who were HIV positive at study entry. ........................................................................... 52

4.4 Estimated bias of the probability HPV infection persists at least $t$ years for scenarios 1-6 .......................................................... 53

4.5 Estimated mean square error (MSE) of the probability HPV infection persists at least $t$ years for scenarios 1-6 ................................................................. 54

4.6 Relative efficiencies (REs) comparing the variance of the SSPLA estimates to the variance of the IO estimates for scenarios 1-6 ................................. 55

5.1 Semi-Markov maximum likelihood estimators (MLEs) of HPV type-16 related persistence from a model assuming transition probabilities for leaving state 1 are semi-Markov and type 52 differs from types 16, 31, 33, 35, and 58 for women in HERS cohort who were HIV positive and HPV negative at study entry. ................................................................. 75

5.2 Markov maximum likelihood estimators (MLEs) of HPV type-16 related persistence from a model assuming transition probabilities for leaving state 1 are Markov and different among types 16, 31, 33, 35, 52 and 58 for women in HERS cohort who were HIV positive at study entry. ................................................................. 76
5.3 Semi-Markov maximum likelihood estimators (MLEs) of HPV type-18 related persistence from a model assuming transition probabilities for leaving state 1 are semi-Markov and the same for types 18, 39, 45, 59 and 68 for women in HERS cohort who were HIV positive and HPV negative at study entry. . . . 

5.4 Semi-Markov maximum likelihood estimators (MLEs) of HPV type-18 related persistence from a model assuming transition probabilities for leaving state 1 are semi-Markov and type 59 differs from types 18, 39, 45 and 68 for women in HERS cohort who were HIV positive at study entry. . . . . . . . . . . . . . 78

5.5 Empirical estimator and semi-Markov maximum likelihood estimators (MLEs) with and without random effects of HPV type-16 persistence among women in HERS cohort who were HIV positive and HPV type 16 negative at study entry. Mixed effects model 5 was used. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79

5.6 Semi-Markov maximum likelihood estimators (MLEs) of HPV persistence from a model assuming transition probabilities leaving state 1 are semi-Markov and types 6, 11, 16 and 18 are different for women in the placebo arm of a vaccine clinical trial (Protocol 005) who were HPV negative for types 6, 11, 16 and 18 at study entry. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80

5.7 Semi-Markov maximum likelihood estimators (MLEs) of HPV persistence from a model assuming transition probabilities leaving state 1 are semi-Markov and types 16 and 18 are the same for women in the placebo arm of a vaccine clinical trial (Protocol 012) who were HPV negative for both types at study entry. . . . 81
Chapter 1

INTRODUCTION

Defining and estimating HPV persistence has important public health implications since HPV persistence is positively associated with the progression of cervical neoplasia (Koshiol et al. (2008); Castle (2008)). In the literature, HPV panel data has been analyzed in many ways through survival analysis methods and multi-state modeling (Koshiol et al. (2006a); Trottier et al. (2008); Liaw et al. (1999); Molano et al. (2003); Xi et al. (2002); Woodman et al. (2001); Kang and Lagakos (2004, 2007); Ho et al. (1998); Franco et al. (1999); Winer et al. (2003)). The most common multi-state model used to analyze longitudinal multi-state panel data relies on the Markov assumption. Though this assumption is well studied and helpful in fitting multi-state models, it may be unrealistic in complex clinical settings. In these settings, it may be more appropriate to use a weaker semi-Markov assumption since this assumption does not depend on time spent in the current state. Continuous-time semi-Markov multi-state models have been proposed for the analysis of panel data from HPV studies but these methods require specification of guarantee times, parametric assumptions, and do not allow for length-biased sampling (Kang and Lagakos (2004, 2007)). Further development of semi-Markov discrete-time multi-state models are needed to account for length bias without requiring such conditions or assumptions. We propose to develop semi-Markov discrete-time multi-state models that can be used to study HPV persistence. The outline of the remaining sections of this thesis are as follows.

Chapter 2 includes a review of the literature. Here we present information concerning the
following: HPV persistence and its role in cervical cancer, definitions of persistence, HPV panel data, existing estimators of persistence, and discrete-time stochastic models.

In Chapter 3 we consider estimators of HPV persistence using panel data on individuals that do not have HPV at study enrollment (i.e., incident infections). We propose a new maximum-likelihood estimator (MLE) of HPV persistence using a semi-Markov two-state discrete-time model for incident infections only. Simulation studies are conducted to compare the proposed MLE and empirical estimators (EEs) currently employed in the HPV literature. We present the MLE and EEs for HPV persistence of incident HPV-16, 18, and 53 infections, separately, in the HERS data (Ahdieh et al. (2001)). We also extend these methods to account for reinfection.

In Chapter 3 we consider incident infections only; however, it is possible that an individual can have HPV at study enrollment. In this case the initiation of the infection is unobserved. Current estimators in the HPV literature exclude these infections altogether or assume the infection started at study enrollment. In Chapter 4, we extend the estimator developed in Chapter 3 using panel data on all individuals including those that have HPV at study enrollment (i.e., prevalent infections). We propose a MLE of HPV persistence using a semi-Markov discrete-time model for prevalent infections. We present the estimates for HPV persistence of any HPV type (i.e., HPV types 6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 45, 51, 52, 53, 54, 55, 56, 58, 59, 66, 68, 73) in the HERS data. A simulation study to confirm the HERS results is conducted to compare the proposed MLE and current estimators. Simulation studies with varying acquisition and clearance assumptions are conducted to examine the bias and precision of the proposed MLE (including prevalent infections) and current MLEs (excluding prevalent infections).

In Chapter 3 we consider type-specific HPV incident infections, separately. It is possible that a subject can have multiple infections (i.e., clustered data) during a study. In Chapter 4, we consider any HPV infection by consolidating an individual's results from multiple types
into a composite outcome. In Chapter 5, we consider incident and prevalent infections from individuals allowing for the possibility that an individual has multiple infections during the study. We assess the differences in persistence among multiple HPV type infections by extending the methods developed in Chapters 3 and 4 to two-state mixed effects models which appropriately handles clustered data. We present the MLE for HPV persistence of high risk types from the HERS data and from phase IIa and III clinical trials of a HPV vaccine. A simulation study is conducted to assess the bias of the MLE.
Chapter 2

BACKGROUND

2.1 The Role of HPV Persistence in Cervical Cancer

Clinical HPV infections are the most common sexually transmitted infections (STIs) reported today, with detection rates of 5%-40% in reproductive women (Franco, Duarte-Franco and Ferenczy (2001)). Carcinogenic HPV infection is a necessary but not a sufficient cause of ICC. There are over 100 HPV DNA types; however, women who are positive for a given HPV DNA type may eventually develop a high grade lesion (Franco et al. (1999)). It has been reported that the risk of CIN is proportional to the number of specimens testing positive for HPV (Ho et al. (1995)). It has also been well established that persistent HPV infection is considered to drive the progression of cervical neoplasia to cervical cancer (Koshiol et al. (2008); Muñoz et al. (1992); Schiffman et al. (1993); Moscicki et al. (2006)), the second most frequently occurring cancer in women worldwide. It is also considered that type-specific persistence is important in the development of cervical carcinogenesis (Koshiol et al. (2008)). The relative risks (RRs) for the association between HPV infection and cervical neoplasia are of high magnitude (RR=20-70 range). Thus, defining and estimating persistence is important in studying the natural disease history of cervical cancer.

The consistently strong positive association between HPV persistence and high-grade cervical intraepithelial neoplasia (CIN 2/3), or pre-cancerous lesions of grades 2 or 3, under various persistence definitions suggests that sequential HPV testing could enhance the identification
of women who are at the highest risk of cervical cancer (Koshiol et al. (2008); Castle (2008)). This association also emphasizes the value of HPV persistence as a clinical marker and endpoint in clinical trials (Koshiol et al. (2008)). In the future, measuring persistence of HPV infection may optimize screening for cervical cancer by increasing sensitivity while maintaining comparable specificity to Pap smear testing. This means essentially that we could detect high grade lesions and potential cervical cancer cases that may otherwise go undetected. Therefore, the future formulation of policies for inclusion of HPV testing in cervical cancer screening and prevention depends on clearly defining persistence, and developing a consensus definition can usefully inform clinical practice for future cervical cancer screening programs.

2.2 Defining Persistence

Despite its importance for the development of CIN and ICC in women, the term “persistence” has often been loosely defined in the HPV literature. Many approaches have been used to define HPV persistence (Woodman and Young (2007)). The differences in HPV persistence definitions are the number of positive testing intervals, the time between HPV tests, the total length of follow-up, and whether HPV infections are present at study enrollment. The duration of persistence of an infection is usually loosely defined as the time from first positive test until a subsequent negative test. For example, if a woman is tested at five time points cataloged as (1,2,3,4,5) and has results (01100) where 0 denotes HPV negative and 1 denotes HPV positive, then the estimated duration of infection would be two units of time, since the first positive test was at the second time point and the first subsequent negative test was at the fourth. However, there is no consensus regarding what exactly constitutes a “persistent” HPV infection. Most studies define persistence as two or more HPV positive visits, three or more HPV positive visits, or time to clearance (Koshiol et al. (2008)). In these definitions, consecutive HPV positive visits are generally required, although single intermittent HPV-negative visits may be allowed (Koshiol et al. (2008)). For example, some authors define persistence as time until a confirmed negative test, where a confirmed negative test is the first of two consecutive negative visits (Koshiol et al. (2006a)). Others define persistence
as testing positive for the majority of the follow-up visits or as testing positive followed by three or four negative test results (Moscicki et al. (1998)). The most commonly used formal definition of HPV persistence is HPV positivity at a minimum of two consecutive follow-up visits with a median interval of six months apart (Rostich et al. (2011)). Ultimately, however, heterogeneity across studies in the time between scheduled follow-up and the many unknowns regarding the natural history of HPV complicate identification of a universally acceptable interval. Thus, the significance of being positive at two consecutive time points becomes blurred, as does the distinction between “persistence” and “transient” infection (Baseman and Koutsky (2005)).

2.3 Panel Data

Typically, epidemiological natural history studies catalog a subject’s transition between a set of well-defined states over the course of time. These studies produce incomplete longitudinal data since the states are observed only at fixed clinic visits. This type of data is often referred to as “panel data,” which is data observed intermittently at a sequence of discrete time points; however, the exact times of transition between states are unobservable (Gentleman et al. (1994); Kalbfleisch and Lawless (1985)). For example, HPV data is collected as a series of HPV results that are either positive (1) or negative (0) as measured by polymerase chain reaction (PCR) or hybrid capture II testing longitudinally at fixed time points. Empirically, a researcher will record the values of these HPV results at each time point. Study subjects can then occupy, one at a time, a set of discrete states (0=HPV negative or 1=HPV positive) in which they can transition (i.e., progress/regress), staying a random amount of time in each state before transitioning to the next. For example, Trottier et al. (2008) followed a cohort of women for five years, with HPV tests every four months for the first year and every six months thereafter. Koshiol et al. (2006a) describe data from participants in the HERS cohort where women were tested for HPV every six months.
Panel data pose many statistical challenges in the context of HPV persistence. Usually, longitudinal data on HPV in a cohort of subjects may be incomplete in several ways (Gentleman et al. (1994)). First, subjects may have missing data. HPV test results may be missing for many reasons: 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. As a result, the HPV status of an individual may be unobservable at some scheduled study visits. If handled inappropriately, misleading inference can occur. For example, we may estimate persistence by carrying the last observation forward; however, doing so may result in overestimating the true duration of infection. Second, it is not possible to observe the duration of infection of individuals who are HPV positive at study enrollment, since the time of the origination of the infection is typically unknown. Third, duration of infection is never observed precisely but are known only to have occurred between two visits; instead, these durations gather around scheduled time points.

2.4 Existing Estimators of Persistence

Many different approaches have been used to estimate persistence (Rostich et al. (2011)). Current analytical approaches typically employed in estimating time of persistence entail using standard survival analysis methods, which we call “empirical estimators” (EEs). The HPV literature contains many examples of EEs. These methods often explicitly or implicitly assume that an individual is HPV positive when she (or he) misses a study visit following a visit where she (or he) tested HPV positive. Intuitively, such an assumption will lead to over-estimation of the duration of infection. For example, Trottier et al. (2008) ignore missing test results, such that missing HPV results between two positive tests are essentially assumed to be positive. A similar approach is used by Koshiol et al. (2006a) when estimating HPV persistence. There are still more examples of EEs that make similar assumptions about missing results (Liaw et al. (1999); Molano et al. (2003); Xi et al. (2002); Woodman et al. (2001)).
In the presence of right-censoring, EEs have been employed to handle missing data caused by dropout when estimating persistence (Molano et al. (2003); Ho et al. (1998); Moscicki et al. (1998); Trottier and Franco (2006); Ahdieh et al. (2000); Liaw et al. (1999); Hildesheim et al. (1994); Kotloff et al. (1998); Woodman et al. (2001); Hogewoning et al. (2003)). For example, the cumulative probabilities of acquiring and clearing HPV infection are often estimated using the Kaplan-Meier method (Ho et al. (1998); Moscicki et al. (1998); Trottier and Franco (2006)). In studying the clearance of HPV in Colombian women, Molano et al. (2003) used Cox regression methods for interval-censored survival data as described by Pan (2000). EEs, like the Kaplan-Meier estimator, have certain drawbacks, however. First, the methods do not allow for intermittent missing data unless specific models for interval censoring are employed. Second, EEs of the average duration of infection do not allow for the possibility of repeat infections of the same HPV type (Molano et al. (2003); Ahdieh et al. (2000)). For example, an HPV-type 16 infected individual can clear infection only for it to reoccur later. In this situation, HPV-type infection is commonly used as the unit of analysis instead of the individual. This inherently assumes that each individual’s infection is independent of one another which is a very strong and potentially dubious assumption. Third, EEs may overestimate persistence in the presence of multiple infections of different types simultaneously if the average duration of infection with at least one type-specific HPV infection is the outcome of interest (Ho et al. (1998)). This overestimation may occur if individuals have multiple infections that overlap study visits.

EEs do not provide a natural way for combining data on HPV positive individuals at study enrollment (i.e., prevalent infections) and HPV negative individuals at study enrollment that subsequently become HPV positive after enrollment (i.e., incident infections). One commonly used approach is to simply combine all infections without making the distinction between prevalent and incident infections (Koshiol et al. (2008)). For example, Koshiol et al. (2006a) analyzed time of persistence of HPV infection by combining the results of both prevalent and incident HPV infections. They used the start date of infection as the first HPV positive visit,
which essentially assumes that a woman with a prevalent infection becomes infected just prior to study enrollment. Alternatively, other methods for estimating persistence exclude women with prevalent infection at enrollment (Trottier et al. (2008)). Such exclusions can lead to lost information that may be important in studying the disease process, especially in settings where pre-disease states can be transient and recurring. Rostich et al. (2011) provide a detailed review of HPV studies that excluded prevalent infections as well as mixed prevalent and incident infections for analyses.

Trottier et al. (2008) showed that prevalent infections take longer to clear than incident infections, probably because the prevalent infections tend to over-represent the most “severe” infections (e.g., those with a high viral load). As the duration of infection increases, the chance of detection as a prevalent case increases. This phenomenon of observation selection with probability proportional to the duration of infection is known as length-biased sampling and has been well studied in many research areas, including epidemiology, when studying survival data for prevalent disease cohorts. Biased sampling can be detrimental to the process of inference on population parameters if not addressed properly. There are many statistical methods developed to handle length-biased sampling (Wang (1998); Wang, Brookmeyer and Jewell (1993)). To our knowledge, none of the persistence estimators employed in the HPV literature utilize methods to adjust for length bias (Ahdieh et al. (2001, 2000); Evander et al. (1995); Liaw et al. (1999); Molano et al. (2003); Bory et al. (2002); Kotloff et al. (1998)).

2.5 Stochastic Models

Another alternative to analyzing panel data is stochastic multi-state modeling (Commenges (1999); Hougaard (1999); Andersen and Keiding (2002); Commenges (2002)). Multi-state models can be employed to provide a comprehensive view of a disease process, allow estimation of proportions of individuals who will be in the various states at some time in the future, and make more efficient use of incomplete information when only fairly short portions of an individual’s disease histories are available (Gentleman et al. (1994)). Multi-state
models have wide applications in biomedical and health research, including HPV (Kang and Lagakos (2004, 2007); Ho et al. (1998); Franco et al. (1999); Koshiol et al. (2006a); Winer et al. (2003)), HIV/AIDS (Kim, DeGruttola and Lagakos (1993); Gentleman et al. (1994); Satten and Longini (1996); Aalen et al. (1997); Satten and Sternberg (1999); Sternberg and Satten (1999); Joly and Commenges (1999); Chen and Tien (2004); Foucher et al. (2005); Healy, De Gruttola and Pagano (2007); Mathieu et al. (2007); Biase et al. (2007)), and cancer (Dinse and Lagakos (1980); Kay (1986); Wanek et al. (1994); Pérez-Ocón, Ruiz-Castro and Gámiz-Pérez (2001); Cook and Lawless (2002); Putter et al. (2006)).

Multi-state models have been well studied under a Markov assumption; however, these models make the very strong assumption that future transitions between states depend only upon the current state, independent of time. This assumption has been used in both discrete- and continuous-time multi-state models (Andersen and Goodman (1957); Goodman (1958); Aalen et al. (1997); Bartholomew (1983); Wasserman (1980); Kalbfleisch and Lawless (1985); Kay (1986); Keiding and Andersen (1989); Frydman (1992, 1995b); Frydman and Szarek (2009); Wanek, Elashoff and Morton (1993); Wanek et al. (1994); Gentleman et al. (1994); DeGruttola and Lagakos (1989)). The Markov property is very appealing, widely used in practice, and immensely helpful in fitting multi-state models because of its parsimony and simplicity. Multi-state Markov models have also been used to estimate HPV persistence (Ho et al. (1998); Franco et al. (1999); Koshiol et al. (2006a); Winer et al. (2003)).

Despite its popularity, the Markov assumption may not be appropriate in modeling certain processes because the probability of transitioning between states may depend on the elapsed time in the current state. For example, it has been shown that the likelihood that an HPV infection will not clear increases as the infection duration increases (Trottier et al. (2008); Plummer et al. (2007); Ho et al. (1998)). Similarly, the Markov assumption would not account for the strong association between duration of infection and progression to cervical abnormality (Stoler (2000)). When future transitions depend upon the time spent in the current
state, the stochastic process is classified as a semi-Markov process. In the HPV setting, semi-Markov models allow for the possibility that the probability of HPV clearance may depend on how long an individual has been HPV positive or that the likelihood of developing a cervical abnormality may depend on the duration of infection.

Semi-Markov models have been extensively studied in continuous time (Pyke (1961\textit{a},\textit{b}); Moore and Pyke (1968); Lagakos, Sommer and Zelen (1978); Dinse and Lagakos (1980); Mode and Pickens (1988); Lawless and Fong (1999); Joly and Commenges (1999); Andersen, Esbjerg and Sorensen (2000); Escolano et al. (2000); Kang and Lagakos (2004); Chen and Tien (2004); Foucher et al. (2005); Mathieu et al. (2007); Kang and Lagakos (2007)). Typically these methods assume the sample paths are continuously observed. However, it is often the case where study individuals’ states are observed only at discrete time points with no information about the occupied states in between observation times.

Recently, Kang and Lagakos (2007) developed methods for fitting continuous-time semi-Markov multi-state models to panel data. Their methods are 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 make progress towards analyzing multi-state panel data without the usual Markov assumption, their proposed methods require some strong assumptions, which arise from their modeling of time as continuous. First, Kang and Lagakos assume that each individual starts in the same infection-free state. That is, their method can be used only to analyze incident infections. Second, they require specification of “guarantee times” 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. Third, they assume that parametric distributions govern the transition intensity functions out of semi-Markov states.

Discrete-time semi-Markov models (Frydman (1995\textit{a}); Satten and Sternberg (1999); Sternberg and Satten (1999); Satten (1999); Barbu, Boussemart and Limnios (2004); Barbu and
Limnios (2006)) have not received as much attention in the literature as continuous-time semi-
Markov models. For studies with fixed scheduled visits, such as clinical trials, it is natural
to model time as discrete. Discrete-time models can have advantages over continuous-time
models, such as not requiring the specification of guarantee times.

Satten and Sternberg (1999) and Sternberg and Satten (1999) studied nonparametric estima-
tors for discrete-time semi-Markov unidirectional models with varying initial states in HIV
data. Their methods allow individuals to have an unknown initial observation time by mod-
eling the transition from an initial positive state to a negative state by a separate nuisance
c Parameter function (Satten and Sternberg (1999); Sternberg and Satten (1999)). Satten and
Sternberg considered only unidirectional models, which may not be applied to complex dis-
ease processes such as HPV where prior states may be revisited and states may not be visited
sequentially.

Frydman (1995a) studied nonparametric estimators for three-state discrete-time semi-Markov
unidirectional models with application to HIV/AIDS. Although the methods extended the
Markov models developed by DeGruttola and Lagakos (1989) and Frydman (1992; 1995b)
to a more generalizable discrete-time semi-Markov framework by allowing the probability of
transitioning from the HIV positive state to AIDS to depend on the duration of HIV infection.
Like Satten and Sternberg, Frydman also considered only unidirectional models. Frydman’s
methods assume that each individual begins in the HIV negative state; thus, the method
only analyzes incident infections. Lastly, Frydman’s methods make parametric assumptions
about the transition from an HIV positive state to AIDS.

Barbu et al. (2004) and Barbu and Limnios (2006) studied discrete-time multi-state bi-
directional semi-Markov models. Their methods require parametric assumptions, only allow
for incident infections, and do not address the possibility of missing data.
2.6 Summary and Proposed Research

In summary, defining and estimating HPV persistence has important public health implications since HPV persistence is positively associated with the progression of cervical neoplasia. In the literature, HPV panel data has been analyzed in many ways through empirical estimators and multi-state modeling. The most common multi-state model used to analyze longitudinal multi-state panel data relies on the Markov assumption. Though this assumption is well studied and helpful in fitting multi-state models, it may be unrealistic in complex clinical settings. In these settings, it may be more appropriate to use a weaker semi-Markov assumption since this assumption does not depend on time spent in the current state. Continuous-time semi-Markov multi-state models have been proposed for the analysis of panel data from HPV studies but these methods require specification of guarantee times, parametric assumptions, and do not allow for length-biased sampling. Further development of semi-Markov discrete-time multi-state models are needed to account for length bias without requiring such conditions or assumptions. We propose to develop semi-Markov discrete-time multi-state models that can be used to study HPV persistence.
Chapter 3

SEMI-MARKOV MODELING OF HPV PERSISTENCE

3.1 Introduction

3.1.1 Defining and estimating HPV persistence

Persistent HPV infection is considered to drive the progression of cervical neoplasia to cervical cancer (Koshiol et al. (2008); Schiffman et al. (1993)), 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) (Koshiol et al. (2008)). Thus, HPV persistence is important as a clinical marker and endpoint in clinical trials (Koshiol et al. (2008)) and in screening to identify women who are at highest risk of high grade pre-cancerous lesions and cervical cancer (Koshiol et al. (2008); Castle (2008)). Characterizing the persistence of HPV infections is also important in studying the natural disease history of cervical cancer.

Despite its importance, there is no consensus regarding what exactly constitutes a “persistent” infection (Koshiol et al. (2008); Woodman and Young (2007)). 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”) (Gentleman et al. (1994); Kalbfleisch and Lawless (1985)). 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. (Koshiol et al. (2006a)) 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 (Woodman and Young (2007)). 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 (Koshiol et al. (2006a); Rostich et al. (2011); Koshiol et al. (2008); Trottier et al. (2008); Liaw et al. (1999); Woodman et al. (2001)). EEs sometimes exclude individuals with consecutively missing HPV results (Koshiol et al. (2006a,b)). 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 (Trottier et al. (2008)). 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 (2007) 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 (Plummer et al. (2007); Maucort-Boulch et al. (2010)). 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.

3.1.2 Outline

The outline of the remaining sections is as follows. In Section 3.2 an EE of type-specific HPV persistence is illustrated using panel data. In Section 3.3 a maximum-likelihood estimator (MLE) of type-specific HPV persistence is proposed using a semi-Markov two-state discrete-time model. Section 3.4 describes a simulation study comparing the MLE and EE in settings similar to HERS. Section 5 extends the two-state model from Section 3.3 to a more general three-state model. In Section 3.6 the different estimators are applied to data from HERS. Section 3.7 concludes with a discussion.

3.2 Empirical estimator

In this section, we present an illustrative example of an EE motivated by Koshiol et al. (2006a). Suppose we observe five subjects in one of two observable states (0=HPV type negative or 1=HPV type positive) at seven time points $\tau_0, \ldots, \tau_6$. The following could occur:

<table>
<thead>
<tr>
<th>Subject</th>
<th>$\tau_0$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$\tau_3$</th>
<th>$\tau_4$</th>
<th>$\tau_5$</th>
<th>$\tau_6$</th>
<th>$\tilde{T}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>*</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>*</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>*</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

where * represents a missing response. In the last column of the table above $\tilde{T}_i$ denotes an estimate of the duration of HPV type infection for individual $i$. Following Koshiol et al. (2006a), 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 $\tau$ time points equals $n^{-1} \sum_{i=1}^n I[\tilde{T}_i > \tau]$, where $I[\cdot]$ is the usual indicator function which
equals 1 if \( \cdot \) is true and 0 otherwise. The estimated times \( \tilde{T}_i \) 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 3.4 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.

### 3.3 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.

#### 3.3.1 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 \( \tau \). 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 \( Y_i \) before transitioning to \( Y_{i+1} \)). Thus, \( X(\cdot) \) is equivalent to \( \{Y_0,T_1,Y_1,\ldots,Y_i,T_i,\ldots\} \). Assume \( X(\cdot) \) is a time-homogeneous semi-Markov process (Barbu, Boussemart and Limnios (2004); Barbu and Limnios (2006)) such that for \( j \neq k \) and \( i = 1,2,\ldots \)

\[
P\{Y_{i+1} = k, T_{i+1} = t | Y_1, \ldots, Y_i = j; T_1, \ldots, T_i \} = P\{Y_{i+1} = k, T_{i+1} = t | Y_i = j\}, \quad (3.1)
\]

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.,

\[
p_{jj}(t) = 1 - \sum_{j \neq k} p_{jk}(t).
\]
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

\[ P \{ X_{i+1} = k | X_i = j, X_{i-1}, \ldots, X_0 \} = P \{ X_{i+1} = k | X_i = j \}. \] (3.2)

Let \( p_{jk} = P \{ X_{i+1} = k | X_i = j \} \). The Markov property (3.2) 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 (Trottier et al. (2008)). 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 (Kang and Lagakos (2007)), 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 (2007). That is, we assume (3.1) holds for \( j = 1, k = 0 \) and that (3.2) holds for \( j = 0, k = 1 \). Models relaxing this assumption are considered in Section 3.5.

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 \( \mathbf{p} = (p_{01}, p_{10}(1), p_{10}(2), \ldots, p_{10}(n_t - 1), p_{1+}(n_t)) \) 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 $\pi_x(p) = P\{(X_0, X_1, \ldots, X_{n_t}) = x\}$. Then $\pi_x(p)$ is

$$p_{y_0y_1}(t_1)p_{y_1y_2}(t_2)\cdots p_{y_{m-2}y_{m-1}}(t_{m-1})p_{y_m+(t_m+)} = \left\{ \prod_{i=0}^{m-2} p_{y_iy_{i+1}}(t_{i+1}) \right\} p_{y_m+}(t_{m+}).$$ (3.3)

### 3.3.2 Estimands

Below we show that HPV type-specific persistence at any particular time point may be written as a function of $p$. Here the operational definition of persistence discussed in Section 3.1 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 $\phi_j(p)$ 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

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

and

$$\phi_3(p) = P\{X_{(0:5)} = (01100)|X_0 = 0, X_1 = 1\} + P\{X_{(0:5)} = (010100)|X_0 = 0, X_1 = 1\} = p_{10}(3)(1 - p_{01}) + p_{10}(1)^2 p_{01}(1 - p_{01}).$$
respectively. More generally, define for \( x_{(0:j+2)} \in [0,1]^{j+3} \), a vector of length \( j+3 \) zeros and ones,

\[
\eta_{x_{(0:j+2)}} = \begin{cases} 
1 & \text{if } x_1 = x_j = 1; x_i + x_{i+1} > 0 \, \forall \, i = 2, \ldots, j - 2; x_{j+1} = x_{j+2} = 0 \\
0 & \text{otherwise}
\end{cases}
\]  

(3.4)

That is, \( \eta_{x_{(0:j+2)}} \) 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, \( \eta_{x_{(0:j+2)}} \) equals zero. Then for \( j = 1, 2, \ldots \)

\[
\phi_j(p) = \sum_{x_{(0:j+2)} \in [0,1]^{j+3}} \eta_{x_{(0:j+2)}} P \{ X_{(0:j+2)} = x_{(0:j+2)} | X_0 = 0, X_1 = 1 \},
\]  

(3.5)

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

\[
1 - \sum_{i=1}^{j-1} \phi_i(p).
\]  

(3.6)

3.3.3 Inference

Maximum likelihood methods can be used to draw inference regarding the estimands of interest, e.g., (3.6). One challenge in the analysis of longitudinal HPV studies is missing data. That is, \( x \) may not be completely observable for some individuals. Assuming the missing data mechanism is missing at random (MAR) (Little and Rubin (1987)) (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 \( \alpha_{i}x = 1 \) if the path \( x \) is consistent with the observed data for that individual, and 0 otherwise. Revisiting the example in Section 3.2, for subject 1, \( \alpha_{1}(0000100) = 1 \) and \( \alpha_{1}x = 0 \) for all \( x \neq (0000100) \); for subject 5, \( \alpha_{5}(0001100) = \alpha_{5}(0001000) = 1 \) and \( \alpha_{5}x = 0 \) otherwise.
Under MAR the likelihood can be written as

\[
L(p) = \prod_{i} \sum_{x \in [0,1]^{n_t+1}} \alpha_i x \pi_x(p) \tag{3.7}
\]

where the product is over all individuals and the sum is over all possible paths of length \( n_t + 1 \).

Let \( \Omega \) denote the set of \( p \) satisfying the constraints

\[
\Omega = \begin{cases} 
p : 0 \leq p_{01} \leq 1, \\
0 \leq p_{10}(t) \leq 1 \text{ for } t \in \{1, \ldots, n_t - 1\}, \\
0 \leq p_{1+}(n_t) \leq 1, \\
\sum_{t=1}^{n_t-1} p_{10}(t) + p_{1+}(n_t) = 1
\end{cases}
\]

Standard numerical methods for maximizing functions with linear equality and inequality constraints can be used to maximize \( \log L(p) \) over \( \Omega \). 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 \( \log L(p) \) over \( \Omega \).

Confidence intervals (CIs) for inference regarding the transition probabilities (as well as functions thereof such as (3.6)) can be obtained using profile likelihood (Murphy and van der Vaart (2000)). 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 \( \chi^2 \) distribution with \( n_t - 2 \) degrees of freedom.

### 3.4 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. (2006a). 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 3.3 to the HERS data for HPV types 16 (scenario I), 53 (scenario II), and 18 (scenario III). For scenario I, the transition probabilities were \( p = (0.016, 0.653, 0.151, 0.085, 0.0001, 0.0001, 0.028, 0.0001, 0.0001, 0.083) \). Based on equation (3.6), 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, \( p = (0.038, 0.526, 0.225, 0.094, 0.059, 0.0001, 0.014, 0.0001, 0.042, 0.042) \). For scenario III, \( p = (0.015, 0.570, 0.247, 0.076, 0.0001, 0.024, 0.0001, 0.0001, 0.0001, 0.081) \).

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 (\( t = \{0.5, 1.0, \ldots, 3.5\} \)) were evaluated. The MLE was computed based on equation (3.6). Here and in the sequel the EE was evaluated using the Kaplan-Meier estimator as described in Koshiol et al. (2006a). 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 3.1. 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 3.1 indicates approximate nominal coverage.

### 3.5 Extensions

In the semi-Markov model developed in Section 3.3, 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 3.3, 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 \( \chi^2 \) 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 3.3, 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 3.3.2, \( p_{01} \) 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 \( \chi^2 \) 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.

3.6 HIV Epidemiology Research Study (HERS)

In this section we analyze data from the HERS cohort described in Section 3.1.1. 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 (Koshiol et al. (2006a)); HPV types 16 and 18 are the two most common cancer-associated HPV types worldwide (Clifford et al. (2003)). 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 3.3 and the three-state models from Section 3.5 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.

3.7 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 (2007) 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 (Nagelkerke, Chunge and Kinoti (1990); Crespi, Cumberland and Blower (2005); Singer and Cohen (1980)), and only panel data is available on the infection/activation state.
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 $\hat{T}$ used in computing the EE (as described in Sections 3.2 and 3.4).

<table>
<thead>
<tr>
<th>$x = (x_0, x_1, x_2, x_3)$</th>
<th>$\hat{T}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0100)</td>
<td>1</td>
</tr>
<tr>
<td>(0111)</td>
<td>3+</td>
</tr>
<tr>
<td>(0101)</td>
<td>3+</td>
</tr>
<tr>
<td>(0110)</td>
<td>2+</td>
</tr>
<tr>
<td>(0010)</td>
<td>1+</td>
</tr>
<tr>
<td>(0011)</td>
<td>2+</td>
</tr>
<tr>
<td>(0001)</td>
<td>1+</td>
</tr>
<tr>
<td>(0000)</td>
<td>*</td>
</tr>
</tbody>
</table>

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 $x = (0111)$, 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 $x = (0101)$ the duration of infection is also at least 3 time units. Note in the last row that $x = (0000)$ does not have an estimated duration of HPV infection $\hat{T}$ 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 3.4, Koshiol et al. (Koshiol et al. (2006a)) 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

\[
\tilde{\phi}(1) = \frac{\sum_i I[x_i = (0100)]}{\sum_i I[x_i \neq (0000)]},
\]

where the summation is over all individuals. The numerator of (3.8) is the number of individuals with an infection of duration 1 time unit and the denominator of (3.8) 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,

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

Clearly the right side of (3.9) does not in general equal \( \phi(1) = p_{10}(1)(1 - p_{01}) \). In fact, the right side of (3.9) 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.
Table 3.1: 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.

<table>
<thead>
<tr>
<th>$t$ (years)</th>
<th>Scenario I</th>
<th>Scenario II</th>
<th>Scenario III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Truth</td>
<td>Coverage</td>
<td>Truth</td>
</tr>
<tr>
<td>0.5</td>
<td>0.36</td>
<td>0.95</td>
<td>0.49</td>
</tr>
<tr>
<td>1.0</td>
<td>0.21</td>
<td>0.95</td>
<td>0.28</td>
</tr>
<tr>
<td>1.5</td>
<td>0.12</td>
<td>0.96</td>
<td>0.18</td>
</tr>
<tr>
<td>2.0</td>
<td>0.12</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>2.5</td>
<td>0.11</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>3.0</td>
<td>0.09</td>
<td>0.96</td>
<td>0.09</td>
</tr>
<tr>
<td>3.5</td>
<td>0.08</td>
<td>0.96</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Table 3.2: 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

<table>
<thead>
<tr>
<th>$t$ (years)</th>
<th>Type 16* (n=49)</th>
<th>Type 53* (n=102)</th>
<th>Type 18** (n=48)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EE</td>
<td>MLE†</td>
<td>EE</td>
</tr>
<tr>
<td>0.5</td>
<td>0.84</td>
<td>0.48 [0.35, 0.61]</td>
<td>0.85</td>
</tr>
<tr>
<td>1.0</td>
<td>0.61</td>
<td>0.34 [0.23, 0.47]</td>
<td>0.58</td>
</tr>
<tr>
<td>1.5</td>
<td>0.44</td>
<td>0.21 [0.12, 0.34]</td>
<td>0.45</td>
</tr>
<tr>
<td>2.0</td>
<td>0.35</td>
<td>0.18 [0.09, 0.30]</td>
<td>0.43</td>
</tr>
<tr>
<td>2.5</td>
<td>0.30</td>
<td>0.15 [0, 0.28]</td>
<td>0.39</td>
</tr>
<tr>
<td>3.0</td>
<td>0.30</td>
<td>0.12 [0, 0.24]</td>
<td>0.31</td>
</tr>
<tr>
<td>3.5</td>
<td>0.20</td>
<td>0.11 [0, 0.24]</td>
<td>0.29</td>
</tr>
</tbody>
</table>

* 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 [ , ]
Figure 3.1: 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 3.2: 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.
Chapter 4
MODELS WITH UNKNOWN INITIAL OBSERVATION TIMES

4.1 Introduction

4.1.1 Estimating persistence

In Chapter 2 we proposed semi-Markov models for incident type-specific infections. However, it is often the case in HPV studies that some women are already infected at study enrollment; these are often called “prevalent” infections. Trottier et al. (2008) showed that prevalent infections take longer to clear than incident infections, probably because the prevalent infections tend to over-represent the most “severe” infections (e.g., those with a high viral load). For women already infected at baseline, estimating the duration or persistence of infection is challenging since the duration of their infection prior to baseline is unknown.

A current method for estimating persistence often excludes women with prevalent infection at enrollment (Trottier et al. (2008); Rostich et al. (2011); Woodman et al. (2001)). We refer to this approach as the “Incidence Only” (IO) approach. Chapter 3 presented several different estimators (i.e., EEs and MLEs) for incident infections only. Exclusions of prevalent infections can lead to loss of information that may be important in studying the disease process. Another method combines both prevalent and incident infections and assumes that prevalent infections developed just prior to study enrollment. We refer to this approach using a semi-Markov estimator as the “Naive semi-Markov Estimator” (NSME). Although NSME
uses all available information it potentially underestimates the duration of infection.

As the duration of infection increases, the chance of detection as a prevalent case increases. This phenomenon of observation selection with probability proportional to the duration of infection is known as length-biased sampling and has been well studied in many research areas, including epidemiology, when studying survival data for prevalent disease cohorts (Wang (1998); Brookmeyer, Gail and Polk (1987); Wang (1996); Wang, Brookmeyer and Jewell (1993)). Biased sampling can lead to misleading inference about population parameters if not addressed properly.

There are many statistical methods developed to handle length-biased sampling (Wang (1998)). Plummer et al. (2007) used discrete-time semi-Markov modeling to handle length biased sampling in HPV; however, their methods used a Bayesian framework. To our knowledge, none of the HPV persistence estimators employed in the HPV literature utilize methods to adjust for length bias in a frequentist framework (Ahdieh et al. (2001, 2000); Evander et al. (1995); Liaw et al. (1999); Molano et al. (2003); Bory et al. (2002); Kotloff et al. (1998); Plummer et al. (2007)).

In studying HIV/AIDS data, Satten and Sternberg (1999) developed discrete-time semi-Markov unidirectional models with varying initial states by introducing a separate nuisance parameter function that accounts for length bias sampling. We refer to this approach as the “Satten and Sternberg/Plummer Approach” (SSPLA). This method adds an additional state for observations first seen an unknown amount of time after beginning the stochastic process. The advantage of this framework is that all the data is used; however, infections at baseline are excluded from the estimation of transition probabilities for infections that develop after baseline.

The issue of unknown initial observation times is easily handled under a Markov assumption since the probability of transitioning from one state to another depends only upon the current
state, independent of time. Incident infections under a Markov assumption for an uninfected state have been studied in Chapter 2; however, an individual in an infected state transitioning to an uninfected state will be analyzed under a semi-Markov assumption where transitions are no longer time independent. When analyzing prevalent infection there is now a need to account for the unknown random amount of time spent in the initial infected state. In this chapter we consider fitting bi-directional discrete-time semi-Markov models using the SS-PLA to data including both incident and prevalent infection using nonparametric frequentist methods that do not require specification of prior distributions or parametric assumptions.

4.1.2 Outline

The outline of the remaining sections is as follows. In Section 4.2 a NSME of HPV persistence is discussed. In Section 4.3 a maximum-likelihood estimator (MLE) of HPV persistence is proposed using an adaptation of Satten and Sternberg/Plummer’s method for a semi-Markov two-state discrete-time model. In Section 4.4 the different estimators (IO, NSME, SSPLA) are applied to high-risk HPV types from HERS. Section 4.5 describes a simulation study comparing the IO, NSME, and SSPLA estimators in settings similar to HERS. Section 4.6 concludes with a discussion.

4.2 Naive semi-Markov estimators (NSME)

In this section, we consider the NSME approach by using the maximum-likelihood estimation methods from Chapter 2. Intuitively, this approach might be more efficient than the IO approach since it uses all of the data (both incident and prevalent infections). However, the NSME approach naively treats prevalent and incident infections the same. Positive HPV results at baseline are assumed to be negative just prior to study entry. This assumption ignores the fact that prevalent infections have been in an infected state for an unknown amount of time.

Suppose $X(\cdot) = \{X_\tau : \tau \in \{0, 1, 2, \ldots\}\}$ denotes a discrete-time stochastic process with
state space \( S = \{0, 1\} \) where 0 denotes being HPV type negative and 1 denotes HPV type positive. Assume as in Chapter 3 that \( X_0 = 0 \). However, suppose the process is not observed until some unknown random time \( Z \in \{1, 2, \ldots\} \). Assume the state prior to study entry is 0 (i.e., \( X_{z-1} = 0 \)). Also assume that states 0 and 1 are Markov and semi-Markov, respectively. Under these assumptions, the observable process \( X(\cdot) \) is now characterized by the \((n_t + 2)\)-dimensional vector

\[
p = (p_{01}, p_{10}(1), p_{10}(2), \ldots, p_{10}(n_t - 1), p_{10}(n_t), p_{1+}(n_t + 1)).
\]

Then the likelihood development and inferential procedures for the two-state model are analogous to those in Chapter 2, except there is one additional parameter to estimate, namely the probability of transition from state 1 to state 0 at some time beyond the last study visit, denoted by \( p_{1+}(n_t + 1) \).

### 4.3 Satten and Sternberg-Plummer approach (SSPLA)

As an alternative to the IO and NSME approaches, an approach that accounts for length bias follows.

Let \( Y_j \) denote the \( j \)-th observed state visited after study entry by the stochastic process. Specifically, let \( Y_0 \) denote the observed state visited by the stochastic process at study entry (i.e., \( Y_0 = X_Z \)). If \( Y_0 = 0 \), incident infections (infections occurring after study enrollment) under a Markov assumption are analyzed. As mentioned in Chapter 2 a semi-Markov model has been proposed in analyzing incident infections.

To analyze prevalent infections assume infections are reported at study enrollment (i.e., \( Y_0 = 1 \)). These types of infections have unknown initial observation times (i.e., the origination of the infection is typically unknown). Let \( m \) denote the number of observed states visited by the process and \( n_t \) the number of possible observed time points after study enrollment. Let \( U \) denote the (unknown) amount of time spent in state \( Y_0 = 1 \) prior to time
$Z$, and $T_j$ denote the $j$-th observed sojourn time where $T_j \in \{1, 2, \ldots\} \text{ and } j = 1, 2, \ldots, m$.

Let $V$ denote the total (unknown and observed) sojourn time in $Y_0$ (e.g., $T_1 + U = V$). Let $t_j, u, v, x, y_j$, and $z$ denote the realizations of the random variables $T_j, U, V, X, Y_j$, and $Z$, respectively. Let $x(u) = (x_{z-u}, x_{z-u+1}, \ldots, x_{z+nt})$ denote the value of the process from time of entry into state 1 prior to time $z$ until the end of the observation process. A diagram illustrating the above notation is shown in Figure 4.1.

Assume that $Z$ is independent of the subsequent transition times (i.e., the time at which an individual was first observed does not predict any subsequent transitions). Let $\rho_u = P\{U = u\}$ and $\rho = (\rho_0, \rho_1, \ldots)$. Then the likelihood contribution for an individual $i$ who is initially observed in state $Y_0 = 1$ is given by

$$\ell_i(\rho, p) = \sum_{u=0}^{\infty} \left\{ \rho_u \sum_{x(u) \in [0,1]^{nt+u}} \alpha_{i,x(u)} \pi_{x(u)}(p) \right\} \quad (4.1)$$

where

$$\alpha_{i,x(u)} = \begin{cases} 1 & \text{if } x(u) \text{ is consistent with the observed data for the } i\text{-th individual} \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

and

$$\pi_{x(u)}(p) = p_{y_0y_1}(t_1 + u)p_{y_1y_2}(t_2) \cdots p_{y_{m-1}y_m}(t_m)p_{y_{m+1}}(t_{m+})$$

$$= p_{y_0y_1}(t_1 + u) \left\{ \prod_{i=1}^{m-1} p_{y_iy_{i+1}}(t_{i+1}) \right\} p_{y_{m+}}(t_{m+}) \quad (4.3)$$

where in general

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

for $j \in \{0, 1\}$.

Maximizing the likelihood (4.1) is challenging because $\rho$ is an infinite dimensional parameter.
Following Satten and Sternberg (1999), we consider the following alternative approach. Let 
\( x(z + t_1 : z + n t) = (x_{z+t_1}, x_{z+t_1}, \ldots, x_{z+nt}) \) denote the value of the process from the time of entry into state 0 after time \( z \) until the end of the observation process such that (4.1) can be equivalently written as

\[
\ell_i(\rho, p) = \left\{ \sum_{u=0}^{\infty} \rho_u p_{0u} y_1(t_1 + u) \right\} \left\{ \sum_{x(z+t_1:z+nt)} \alpha_i x(z+t_1:z+nt) \pi(x(z+t_1:z+nt))(p) \right\}
\]

(4.4)

where \( \alpha_i x(z+t_1:z+nt) \) is defined analogously to (4.2). Letting \( p_{20}(t_1) = \sum_{u=0}^{\infty} \rho_u p_{10}(t_1 + u) \) and \( p_{20} = (p_{20}(1), p_{20}(2), \ldots) \), 4.4 can be written as

\[
\ell_i(p_{20}, p) = p_{20}(t_1) \left\{ \sum_{x(z+t_1:z+nt)} \alpha_i x(z+t_1:z+nt) \pi(x(z+t_1:z+nt))(p) \right\}.
\]

(4.5)

Under this parameterization the observable process \( X(\cdot) \) is now characterized by the \((nt + 1)\)-dimensional vector \( p = (p_01, p_{10}(1), p_{10}(2), \ldots, p_{10}(nt-1), p_{1+}(nt)) \) and the \((nt + 1)\)-dimensional vector \( p_{20} = (p_{20}(1), \ldots, p_{20}(nt), p_{2+}(nt+1)) \) where

\[
p_{2+}(t) = 1 - \sum_{i=1}^{t-1} p_{20}(i).
\]

Note \( p_{20}(t_1) \) is the probability of transitioning to a negative state after observed sojourn time \( t_1 \) in a positive initial state. Because the initial state is positive, the calculation of \( p_{20}(t_1) \) involves summing over all unknown sojourn times. This differs from \( p_{10}(j) \) in Chapter 2 which does not require a summation because all initiation times are observed.

Instead of summing over all individuals and maximizing (4.1) as a function of \( \rho \) and \( p \), (4.5) is summed over all individuals and maximized as a function of \( p_{20} \) and \( p \), ignoring the
relationship between $p_{20}$ and $p_{10}$. Let $p_{20}$ satisfy the constraints

$$\tilde{\Omega} = \begin{cases} 0 \leq p_{20}(t) \leq 1 & \text{for } t \in \{1, \ldots, n_t\}, \\ 0 \leq p_{2+}(n_t + 1) \leq 1, \\ \sum_{t=1}^{n_t} p_{20}(t) + p_{2+}(n_t + 1) = 1 \end{cases}$$

The MLE of $p$ is obtained by summing over all individuals and maximizing the log-likelihood, $\log \ell_i(p_{20}, p)$, over $p \in \Omega$ and $p_{20} \in \tilde{\Omega}$ where $\Omega$ is defined in equation (8) of Chapter 2. The MLE of $p$ in turn gives rise to the MLE of functions of $p$ (e.g., $\phi_j(p)$) due to the invariance property of maximum likelihood (Casella and Berger (2002)).

Note the likelihood (4.5) factors $p_{20}$ and $p$ into separate pieces. Thus, we can simply ignore $p_{20}$ and yield the same persistence estimates if we did not ignore $p_{20}$. Ignoring $p_{20}$ provides a computationally easier way to estimate persistence. On the other hand, jointly maximizing (4.5) with respect to $p_{20}$ and $p$ provides inference about the observed time of persistence in prevalent infections.

Conceptually, we can think of (4.5) as arising from a three-state model where we differentiate HPV positive at study entry (state 2) from HPV positive following HPV negative after study entry (state 1), with $p_{20}(t_1)$ denoting the probability of transitioning from state 2 (HPV positive at study entry) to state 0 (HPV negative) after remaining in state 2 for $t_1$ amount of time. Graphically, we can represent this new model as presented in Figure 4.2 which shows that a subject can be observed to move from state 0 (HPV negative) to state 1 or vice versa; however, a subject in state 2 can transition only to state 0. For this model only persons whose initial observation times are known are started in state 0. All other individuals first seen in a positive state are taken into the process in state 2 using their time of study enrollment as the ‘known’ initial observation time. Thus, all individuals may be considered to have known initial observation times, which can have computational advantages (Satten and Sternberg (1999); Sternberg and Satten (1999); Plummer et al. (2007)). Unfortunately, this model is more complex than the two state model presented in Chapter 2, and the methods require the
estimation of several nuisance parameters, $\mathbf{p}_{20}$, which may be inefficient since some data are not used to estimate parameters of interest.

4.4 HIV Epidemiology Research Study (HERS)

In this section we analyze data from the HERS cohort which included 571 women who were HIV seropositive at baseline and had two or more study visits. Our interest is in estimating the persistence of any HPV type (i.e., HPV types 6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 45, 51, 52, 53, 54, 55, 56, 58, 59, 66, 68, 73) for individuals with varying initial states (i.e., both prevalent and incident infections). For the purposes of this paper, persistence is defined as the time between the date of occurrence of the first HPV positive sample of any type(s) and the date of the first occurrence of two consecutive visits negative for the same HPV type(s) (Schiffman, Wheeler and Castle (2002)). For instance, a woman may become infected with types 16 and 18 and both infections are observed for the first time at the same visit. From this time of first occurrence, the woman will remain persistent until both 16 and 18 infections are cleared (i.e., observed to have confirmed type 16 and type 18 negative results). This same woman, may have other types develop at a visit after types 16 and 18 were first observed; however, we will not take these new infections into account. Our interest lies in documenting persistence for the types involved in the first occurrence of infection. Of the 571 women who were HIV seropositive at baseline and had two or more study visits, 202 (35.4%) women were positive for any of the high-risk types listed above at baseline.

Persistence was estimated using the IO, NSME, and SSPLA approaches. 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. Results are shown in Table 4.1. There were a total of 425 infections and 202 (47.5%) were prevalent infections. The estimates of any HPV persistence from IO, NSME, and SSPLA approaches are plotted in Figure 4.3. The estimated probability of persisting at least six months (0.5 years) was 0.62 using the NSME compared to 0.55 or 0.53 using the IO or SSPLA MLE, respectively. The profile likelihood-based 95% confidence
intervals for the SSPLA MLEs were similar to the 95% confidence intervals for the IO MLEs.

4.5 Simulation studies

4.5.1 Simulation to confirm HERS results

A simulation study was conducted to assess the bias of the three MLEs of persistence (approaches 1-3) described above. For the simulation study, persistence is defined as the time between the date of occurrence of the first HPV positive sample and the date of the first occurrence of two consecutive visits recording negative samples. Data sets, each with a sample size of 570 individuals, were randomly generated to be similar to the HERS data. For each individual, data for 60 visits was generated; however, only the last 10 visits generated were used in the analysis. The total number of possible observed time points after study entry was \( n_t = 9 \). The stochastic process transitioning from state 0 (HPV negative) was Markov and transitioning from state 1 (HPV positive) was semi-Markov. 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, 0.08, and 0.09 for study visits 2-9, respectively.

Simulations were based on fitting the two-state model from Section 4.4 to the HERS data for any HPV infections. The transition probabilities were \( p = (0.11, 0.53, 0.14, 0.08, 0.08, 0.01, 0.02, 0.05, 0.000, 0.09) \). This combination of probabilities using random walks ensures a prevalence of 30%.

1000 data sets were generated. For each simulated data set, the MLEs for the IO, NSME, and SSPLA approaches of the probability of any HPV type infection persisting at least \( t \) years (\( t \in \{0.5, 1.0, \ldots, 3.5\} \)) were evaluated. The MLEs were computed based on equation (6) of Chapter 2.

The average estimates for the 3 methods over the 1000 simulations were calculated and the
bias for the 3 methods are presented in Table 4.2. As expected, the NSME tended to over-estimate the probability of HPV persistence. On the other hand, the IO and SSPLA MLEs were approximately unbiased.

The mean square errors (MSEs) for the 3 methods over the 1000 simulations as well as the relative efficiencies (REs) comparing the variance of the SSA estimates to the variance of the IO estimates are presented in Table 4.2. A RE greater than 1 means that the IO estimate is more efficient; in contrast, an RE less than 1 means that the SSPLA estimate is more efficient. As expected, the SSPLA method was more efficient than the IO method.

For each simulated data set, 95% profile likelihood based CIs associated with the SSPLA MLEs were calculated. Empirical coverage 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 indicated approximate nominal coverage (i.e., 94%-96%) for each HPV persistence estimate at least $j = 1, \ldots, 7$ units of time (results not shown).

### 4.5.2 Simulations with varying acquisition and clearance assumptions

We have shown above that the NSME MLEs are biased. We now further examine the bias and precision of the SSPLA and IO MLEs of persistence described above. The stochastic process transitioning from state 0 (HPV negative) was Markov and transitioning from state 1 (HPV positive) was also Markov. The pattern of missingness detailed above was assumed for these studies.

Simulations were based on fitting the two-state model from Section 4.4. There were six scenarios that are combinations of probabilities for acquisition and clearance. Scenarios 1-3 assume the acquisition probability is 0.05 (similar to type-specific acquisition) and the clearance probabilities are 0.25, 0.50, and 0.80, respectively. Scenarios 4-6 assume the acquisition probability is 0.35 (similar to any high-risk type acquisition) and the clearance probabilities
are 0.25, 0.50, and 0.80, respectively.

The average estimates for the 2 methods over the 1000 simulations were calculated and the bias and MSEs are presented in Figures 4.4 and 4.5. Under all six scenarios the IO and SSPLA MLEs were approximately unbiased.

The relative efficiencies (REs) comparing the variance of the SSPLA estimates to the variance of the IO estimates are presented in Figure 4.6. Under all six scenarios the SSPLA method was more efficient than the IO method since the REs were less than 1. The efficiencies were more pronounced when the acquisition probability increased from 0.05 to 0.35.

4.6 Discussion

In this Chapter we proposed an estimator that accounts for length bias when analyzing prevalent infections. We compared this estimator (SSPLA) with two estimators: (1) IO which excludes prevalent infections altogether and (2) NSME which ignores length bias making no distinction between incident and prevalent infections. The NSME MLEs were biased; the SSPLA and IO MLEs were approximately unbiased. The proposed SSPLA and IO MLEs had similar confidence intervals when analyzing the HERS data. Simulation studies showed that the efficiency of the SSPLA MLEs increased (IO MLEs decreased) as prevalence increased.

In the HPV setting where type-specific prevalence rates are low (similar to simulation Scenarios 1-3), the IO method is approximately unbiased with slightly less precision than the SSPLA. This is because there are not many infections that are being thrown out of the IO analysis. The IO method is less complex and easily understood; therefore, it is our recommendation that the IO method is an acceptable method if the prevalence rates are low.

In contrast, if the prevalence rates are high (similar to simulation Scenarios 4-6), the SSPLA
method much more precise than the IO method. This is because the IO analysis is throwing out much more of the infections. The SSPLA method, however, is more computationally complex. It also ignores the relationship between the probability of being in an initial state an unknown amount of time and the transition probabilities. More research is needed in addressing this relationship.

We used a composite endpoint which treats an individual with many types and an individual with a single type infection similarly; however, the burden of disease may be markedly different for both individuals. Furthermore, there may be interest in comparing the persistence across types. Future research is needed to account for multiple infections; this topic is discussed in the next Chapter.
Figure 4.1: Random initial observation of a two-state stochastic process
Table 4.1: Estimated probability HPV persists at least $t$ years for the Incidence Only (IO), Naive (NSME), and SSPLA semi-Markov maximum likelihood estimators based on the 571 women in HERS cohort who were HIV positive at study entry with at least 2 study visits

<table>
<thead>
<tr>
<th>$t$ (years)</th>
<th>NSME (n=425 total infect.)</th>
<th>IO (n=223 incid. only)</th>
<th>SSPLA (n=425 total infect.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.62 [0.58, 0.66]†</td>
<td>0.55 [0.49, 0.61]</td>
<td>0.53 [0.47, 0.58]</td>
</tr>
<tr>
<td>1.0</td>
<td>0.51 [0.47, 0.55]</td>
<td>0.41 [0.35, 0.47]</td>
<td>0.40 [0.35, 0.46]</td>
</tr>
<tr>
<td>1.5</td>
<td>0.41 [0.36, 0.45]</td>
<td>0.30 [0.24, 0.36]</td>
<td>0.30 [0.25, 0.36]</td>
</tr>
<tr>
<td>2.0</td>
<td>0.34 [0.30, 0.38]</td>
<td>0.23 [0.17, 0.28]</td>
<td>0.22 [0.18, 0.27]</td>
</tr>
<tr>
<td>2.5</td>
<td>0.29 [0.25, 0.34]</td>
<td>0.20 [0.15, 0.26]</td>
<td>0.20 [0.16, 0.26]</td>
</tr>
<tr>
<td>3.0</td>
<td>0.25 [0.21, 0.29]</td>
<td>0.17 [0.12, 0.23]</td>
<td>0.17 [0.13, 0.22]</td>
</tr>
<tr>
<td>3.5</td>
<td>0.20 [0.16, 0.24]</td>
<td>0.12 [0.07, 0.18]</td>
<td>0.12 [0.08, 0.18]</td>
</tr>
</tbody>
</table>

† Profile likelihood-based 95% confidence intervals in [ , ]
Table 4.2: Estimated bias, MSE, and RE of the probability HPV infection persists at least \( t \) years from the simulation study described in section 4.5.

<table>
<thead>
<tr>
<th>( t ) (years)</th>
<th>Bias ( \times 10^{-3} )</th>
<th>MSE ( \times 10^{-5} )</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>105.04</td>
<td>1103.40</td>
<td>0.71</td>
</tr>
<tr>
<td>1.0</td>
<td>69.27</td>
<td>479.90</td>
<td>0.74</td>
</tr>
<tr>
<td>1.5</td>
<td>54.04</td>
<td>292.06</td>
<td>0.74</td>
</tr>
<tr>
<td>2.0</td>
<td>39.28</td>
<td>154.31</td>
<td>0.75</td>
</tr>
<tr>
<td>2.5</td>
<td>3.07</td>
<td>0.98</td>
<td>0.80</td>
</tr>
<tr>
<td>3.0</td>
<td>-31.08</td>
<td>96.66</td>
<td>0.81</td>
</tr>
<tr>
<td>3.5</td>
<td>-36.89</td>
<td>136.13</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Figure 4.2: Model to differentiate HPV positive at study entry (state 2) from HPV positive following HPV negative (state 1)
Figure 4.3: Incidence Only (IO), naive (NSME), and Satten and Sternberg-Plummer adapted (SSPLA) semi-Markov maximum likelihood estimators (MLEs) of any HPV type persistence among women in HERS cohort who were HIV positive at study entry.
Figure 4.4: Estimated bias of the probability HPV infection persists at least t years for scenarios 1-6.
Figure 4.5: Estimated mean square error (MSE) of the probability HPV infection persists at least 4 years for scenarios 1-6.
Figure 4.6: Relative efficiencies (REs) comparing the variance of the SSPLA estimates to the variance of the IO estimates for scenarios 1-6.
Chapter 5

SEMI-MARKOV TWO-STATE MIXED EFFECTS MODELS

5.1 Introduction

Persistent infection of high risk HPV types (i.e., HPV types that are most prevalent in invasive cervical cancer (ICC)) is central to cervical cancer. There are over 100 HPV types and 40 of these infect the genital tract. Subjects participating in HPV epidemiological studies are often followed during a period of time and the statuses (positive or negative) for these HPV types are reported at regularly scheduled visits. It is common for a woman to have concurrent or sequential type infections which may not all have the same persistence patterns. Studying the association between HPV type and persistence of HPV infections may identify women at an increased risk of ICC. The association between type and persistence has been inconsistent as conclusions range from a positive association to no association at all (Koshiol et al. (2006a)). This study design creates observations for an individual that are not independent and identically distributed but rather come in clusters that are correlated. This issue of clustering must be addressed when estimating persistence. The objective of this chapter is to assess differences in persistence (i.e., duration of infection) among different incident HPV type infections while appropriately handling clustering in HIV seropositive women enrolled in HERS.
5.1.1 Clustering

There are several naive approaches to handling clustering. One approach is to simply ignore the correlation of multiple infections within a subject (i.e., assume independence). However, there is evidence to indicate that the life cycles of different HPV types are not independent of each other (Woodman and Young (2007)). The strict assumption of independence among types within an individual does not account for unmeasured host or environmental factors that may create correlations in the incidence of different HPV types (Plummer et al. (2007)). Another approach analyzes each type-specific infection separately (Mitchell et al. (2011)). Even though it is clearly biologically relevant to study HPV positivity in terms of individual types, this approach does not allow for hypothesis testing to compare the persistence of different HPV types. Yet another approach uses a composite endpoint as in Chapter 3; however, this approach treats an individual with many types and an individual with a single type infection the same even though the burden of disease may be markedly different for both individuals.

Cox regression models accounting for within-subject correlation have been used to analyze clustered time to event data. These models have been used to analyze the HERS data (Koshiol et al. (2006a)). This is an improvement upon naive approaches mentioned above as these models account for within-subject correlation; however, there are some limitations. First, an infection can be cleared only to reoccur at a subsequent visit; however, standard survival analysis cannot handle this recurrence of an infection. Second, the duration of each scheduled visit is six months which reflects a discrete-time measurement that is not measured exactly and leads to an interval-censored outcome.

Kong et al. (2010) proposed methods that directly model the likelihood of the arbitrarily censored data (including left, interval and right censoring) while also addressing clustering by using parametric frailty models. These methods were applied to the male circumcision trial conducted in Rakai, Uganda. The advantage of frailty models is that the parametric form
allows for ease of computation. SAS NL MIXED can be used to implement procedures. The disadvantage is that it is hard to perform model diagnostics to check assumptions concerning the parametric form (Kong et al. (2010)).

Alternatively, we propose mixed effects discrete-time semi-Markov models are used to analyze clustered panel data. These models can fully utilize the data, appropriately account for the correlations, and yield interpretable results that can allow comparisons of different types. The persistence patterns of multiple high risk HPV types will be presented and compared using these models. To our knowledge this is the first time type-specific persistence patterns have been compared using mixed models in the HERS data as well as in the two clinical studies.

5.1.2 Outline

The outline of the remaining sections is as follows. In Section 5.2 we present mixed-effects discrete-time semi-Markov models. In Section 5.3 the model is applied to data from HERS and from phase IIa and III clinical trials of a HPV vaccine. Section 5.4 describes a simulation study assessing the bias of the MLE. Section 5.5 concludes with a discussion.

5.2 Mixed effects discrete-time semi-Markov model

In this section we extend the methods developed in Chapters 3 and 4 to allow for an individual to be infected with multiple HPV types simultaneously. A mixed effects modeling approach is employed under the assumption type-specific infections may be correlated within individuals, but are independent between individuals.

5.2.1 Incident infections

In the development below we let $j$ denote the type of HPV infection, and without loss of generality we assume $j \in \{1, 2, 3, \ldots, n_J\}$ where $n_J$ is the number of HPV types considered.
in a particular study. Let $X_j(\cdot) = \{X_j^\tau : \tau \in \{0, 1, 2, \ldots\}\}$ denote a discrete-time stochastic process for the $j$-th HPV type infection with state space $S = \{0, 1\}$ where $X_j^\tau = 0$ denotes being HPV type $j$ negative and $X_j^\tau = 1$ denotes being HPV type $j$ positive at time $\tau$. Assume for now $X_j^0 = 0$, i.e., all individuals are HPV type $j$ negative at time 0; this assumption will be relaxed in Section 5.2.2 below. Let $Y_j^k \in \{0, 1\}$ denote the $k$-th state visited by the stochastic process, and let $T_j^k \in \{1, 2, \ldots\}$ denote the $k$-th sojourn time for the $j$-th HPV type infection (i.e., the amount of time that an individual stays in $Y_j^{k-1}$ before transitioning to $Y_j^k$).

Let $p_{ij}^{10}(t)$ denote the probability that the $i$-th individual transitions to state 0 after sojourn time $t$ in state 1 for HPV type $j$. For the $i$-th individual, the hazard of the type $j$ infection clearing at sojourn time $T_j^k = t$, denoted by $h_{10}^{ij}(t)$, is the probability that the type $j$ infection clears at $T_j^k = t$, given the type $j$ infection has not cleared before $T_j^k = t$, i.e.,

$$h_{10}^{ij}(t) = Pr\left[T_j^k = t| T_j^k \geq t, Y_j^{k-1} = 1\right] = \frac{Pr\left[T_j^k = t| Y_j^{k-1} = 1\right]}{Pr\left[T_j^k \geq t| Y_j^{k-1} = 1\right]} = \frac{p_{ij}^{10}(t)}{1 - \sum_{\tau=1}^{t-1} p_{ij}^{10}(\tau)}.$$

Because state 1 is assumed semi-Markov the hazard $h_{10}^{ij}(t)$ does not depend on $k$ or, more generally, the history of the stochastic process prior to entering state 1.

For the $i$-th individual, the log odds of an HPV type $j$ infection clearing at time $t$ given it has not cleared prior to $t$ may be modeled using a logistic-normal mixed effects model Agresti (2002)

$$\log \left( \frac{h_{10}^{ij}(t)}{1 - h_{10}^{ij}(t)} \right) = \beta_{1j} + \eta_t + \delta_i. \quad (5.1)$$

Under model (5.1), for $j, k \in \{1, 2, \ldots, n_J\}$, $\exp(\beta_{1j} - \beta_{1k})$ is the conditional odds ratio of clearing a type $j$ infection versus a type $k$ infection at any time $t$. Similarly, for $s, t \in \{1, 2, \ldots, n_t-1\}$, $\exp(\eta_s - \eta_t)$ is the conditional odds ratio of clearing an infection at time $s$ versus time $t$ for any type infection. To account for correlation between type-specific infections within the same individual $i$, the model (5.1) includes a random effect parameter $\delta_i$ which is unobserved and assumed to be normally distributed with mean 0 and variance $\sigma^2$.  

59
Note the probability of individual $i$ clearing an infection with HPV type $j$ after one unit of time equals

$$p_{10}^{ij}(1) = h_{10}^{ij}(1) = \logit^{-1}(\beta_{1j} + \eta_1 + \delta_i), \quad (5.2)$$

where $\logit^{-1}(x) = (1 + \exp(-x))^{-1}$. Similarly, the probability of clearing an infection after two units of time equals

$$p_{10}^{ij}(2) = h_{10}^{ij}(2) \left\{ 1 - p_{10}^{ij}(1) \right\}$$

$$= \logit^{-1}(\beta_{1j} + \eta_2 + \delta_i) \left\{ 1 - \logit^{-1}(\beta_{1j} + \eta_1 + \delta_i) \right\} \quad (5.3)$$

and so on. In other words, the transition probabilities can be expressed as functions of the parameters $\beta_1, \eta$, and $\delta_i$ where $\beta_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1n_J})$ and $\eta = (\eta_1, \eta_2, \ldots, \eta_{n_t-1})$. Analogous to $\pi_x(p)$ in Chapter 3, let $\pi_x^{ij}(p_{01}, \beta_{1j}, \eta, \delta_i)$ denote the path probability for type $j$ for individual $i$ under model (5.1). Under MAR the likelihood can be written as

$$L(p_{01}, \beta_1, \eta, \sigma) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{n_J} \sum_{x \in [0,1]^{nt+1}} \alpha_{ijx} \pi_x^{ij}(p_{01}, \beta_{1j}, \eta, \delta_i) \right\} f(\delta_i, \sigma) d\delta_i, \quad (5.4)$$

where $f(\delta_i, \sigma)$ is the density of a normal random variables with mean 0 and variance $\sigma^2$ and $\alpha_{ijx}$ is defined analogously to $\alpha_{ix}$ in Chapter 3. The MLE of $(p_{01}, \beta_1, \eta, \sigma)$ is obtained by maximizing $\log L(p_{01}, \beta_1, \eta, \sigma)$ subject to the constraints $0 \leq p_{01} \leq 1$ and $\sigma \geq 0$. Note there are no constraints placed on the parameters $\beta_{1j}$ for $j = 1, 2, \ldots, n_J$ or $\eta$ for $t = 1, 2, \ldots, n_{t-1}$.

The logit transformation used in model (5.1) ensures that

$$0 \leq p_{10}^{ij}(t) \leq 1 \text{ for } t \in \{1, \ldots, n_t - 1\}, \quad (5.5)$$

That (5.5) holds for $t = 1$ follows from (5.2), for $t = 2$ follows from (5.3), and so forth. Note model (5.1) also ensures that

$$\sum_{\tau=1}^{t} p_{10}^{ij}(\tau) \leq 1. \quad (5.6)$$
for all $t \in \{1, 2, \ldots, n_{t-1}\}$. To see that (5.6) holds, note

$$p_{10}^{ij}(2) = \logit^{-1}(\beta_{1j} + \eta_2 + \delta_i) \left\{1 - p_{10}^{ij}(1) \right\} \leq 1 - p_{10}^{ij}(1)$$

implying $p_{10}^{ij}(1) + p_{10}^{ij}(2) \leq 1$. Similarly,

$$p_{10}^{ij}(3) = \logit^{-1}(\beta_{1j} + \eta_3 + \delta_i) \left\{1 - p_{10}^{ij}(1) - p_{10}^{ij}(2) \right\} \leq 1 - p_{10}^{ij}(1) - p_{10}^{ij}(2)$$

implying $p_{10}^{ij}(1) + p_{10}^{ij}(2) + p_{10}^{ij}(3) \leq 1$ and so forth.

The integral in (5.4) does not have a closed form and therefore must be evaluated numerically. In the results below, Gaussian quadrature (Thisted (1986)) was used to evaluate (5.4). The MLEs in turn give rise to the MLEs of functions of $(p_{01}, \beta_1, \eta, \sigma)$ due to the invariance property of maximum likelihood (Casella and Berger (2002)).

### 5.2.2 Prevalent infections

Now we consider mixed effects discrete-time semi-Markov models that account for length bias sampling by extending methods proposed in Chapter 3.

Suppose the stochastic processes $X^1(\cdot), X^2(\cdot), \ldots, X^{n_J}(\cdot)$ are not observed until some unknown random time $Z \in \{0, 1, 2, \ldots\}$. Following Satten and Sternberg (1999) and Plummer et al. (2007) for the $j$-th HPV type let $Y_{k}^{j} \in \{0, 1, 2\}$ denote the $k$-th observed state visited by the stochastic process. State 0 denotes no infection, state 1 denotes an infection developed after study entry, and state 2 denotes an infection at study entry. Specifically, $Y_0^{j}$ denotes the first observed state visited by the stochastic process (i.e., $Y_0^{j} = X_2^j$). For each individual there are $n_J$ stochastic processes; one process for each of the $J$ type infections. All infections may not be detected at study entry.
Suppose $Y^j_0 = 1$ for the $j$-th HPV type infection to analyze multiple prevalent infections under a semi-Markov assumption. The $j$-th infection has an unknown initial observation time. For the $j$-th HPV type infection let $m^j$ denote the number of observed states visited by the process and $n_t$ the number of possible observed time points after study enrollment. Let $U^j$ denote the (unknown) amount of time spent in state $Y^j_0 = 1$ prior to time $Z$, and $T^j_k \in \{1, 2, \ldots \}$ denote the $k$-th sojourn time for the $j$-th HPV type infection. Let $V^j$ denote the total (unknown and observed) sojourn time in $Y^j_0$ (e.g., $T^j_1 + U^j = V^j$). Let $x^j$, $y^j_k$, and $z$ denote the realizations of the random variables $T^j_k$, $U^j$, $V^j$, $X^j$, $Y^j_k$, and $Z$, respectively. Let $x^j(t)$ denote the value of the process from time of entry into state 1 prior to time $z$ until the end of the observation process. In Chapter 3 a diagram illustrating notation for one infection was shown in Figure 4.1. We now have $j$ of these diagrams representing the process for each infection. Similar assumptions about the independence of $Z$ are made here as in Chapter 3.

Let $p^{ij}_{20}(t)$ denote the probability that the $i$-th individual transitions to state 0 after sojourn time $t$ in state 1 for HPV type $j$ observed at study entry. As noted in Chapter 4, because the initial state is positive, the calculation of $p^{ij}_{20}(t)$ involves summing over all unknown sojourn times. This differs from $p^{ij}_{10}(t)$ in Section 5.2.1 which does not require a summation because all initiation times are observed. For the $i$-th individual, the hazard of the type $j$ prevalent infection clearing at sojourn time $T^j_1 = t_1$, denoted by $h^{ij}_{20}(t_1)$, is the probability that the type $j$ infection clears at $T^j_1 = t_1$, given that the type $j$ infection has not cleared before $T^j_1 = t_1$, i.e.,

$$h^{ij}_{20}(t_1) = Pr \left[ T^j_1 = t_1 \mid T^j_1 \geq t_1, Y^j_0 = 1 \right] = \frac{Pr \left[ T^j_1 = t_1 \mid Y^j_0 = 1 \right]}{Pr \left[ T^j_1 \geq t \mid Y^j_0 = 1 \right]} = \frac{p^{ij}_{20}(t_1)}{1 - \sum_{\tau=1}^{t_1-1} p^{ij}_{20}(\tau)}.$$

As in the case with state 1 in Section 5.2.1, because state 2 is assumed to be semi-Markov $h^{ij}_{20}(t_1)$ does not depend of the history of the stochastic process prior to entering state 2.

For the $i$-th individual, the log odds of an HPV type $j$ infection clearing at time $t_1$ given
it has not cleared prior to $t_1$ may be modeled using a logistic-normal mixed effects model
Agresti (2002).

$$\log \left( \frac{h_{20}^{ij}(t_1)}{1 - h_{20}^{ij}(t_1)} \right) = \beta_{2j} + \eta_t + \delta_i$$

(5.7)

where $\delta_i \sim N(0, \sigma^2)$. For $j, k \in \{1, 2, \ldots, n_J\}$, $\exp(\beta_{2j} - \beta_{2k})$ is the conditional odds ratio
of clearing a type $j$ prevalent infection versus a type $k$ prevalent infection at any time $t$. The
interpretation of $\eta_t$ and $\delta_i$ is the same as presented in Section 5.2.1.

The transition probabilities are now expressed as functions of $\beta_1, \beta_2, \eta,$ and $\delta_i$ where $\beta_2 =
\{\beta_{21}, \beta_{22}, \ldots, \beta_{2n_J}\}$ and $\eta = \{\eta_1, \eta_2, \ldots, \eta_{n_t}\}$. Note that $\eta_{n_t}$ can now be estimated since we
are including prevalent infections. This was not the case in Section 5.2.1.

Under MAR the likelihood can be written as

$$L(p_01, \beta_1, \beta_2, \eta, \sigma) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{n_J} p_{20}^{ij}(t_1) \Lambda(p_01, \beta_{1j}, \eta, \delta_i) \right\} f(\delta_i, \sigma) d\delta_i,$$

(5.8)

where

$$\Lambda(p_01, \beta_{1j}, \eta, \delta_i) = \sum_{x(z+t_1:z+n_t)} \alpha_{ij} x(z+t_1:z+n_t) \pi^{ij}_{x(z+t_1:z+n_t)} (p_01, \beta_{1j}, \eta, \delta_i).$$

Similarly as in (5.2) and (5.3), $p_{20}^{ij}(t_1)$ is expressed as functions of $\beta_2, \eta,$ and $\delta_i$. The MLE of
$(p_01, \beta_1, \beta_2, \eta, \sigma)$ is obtained by maximizing $\log L(p_01, \beta_1, \beta_2, \eta, \sigma)$ subject to the constraints
$0 \leq p_{01} \leq 1$ and $\sigma \geq 0$. Note the log transformation used in model (5.7) ensures that

$$0 \leq p_{20}^{ij}(t) \leq 1 \text{ for } t \in \{1, \ldots, n_t\},$$

(5.9)

and also ensures that

$$\sum_{\tau=1}^{t} p_{20}^{ij}(\tau) \leq 1.$$  

(5.10)

The arguments proving both constraints are similar to those given in section 5.2.1.
5.3 Results

For the analysis of HERS and two clinical studies for a HPV vaccine, infection is treated as the unit of analysis since women could have multiple type HPV infections. Persistence is defined as the time from first HPV type-positive result until the time of the first of two consecutive HPV type-negative results. Infections are classified as “incident” if they were first detected at a post-baseline visit and “prevalent” if they were first detected at baseline.

5.3.1 HIV Epidemiology Research Study (HERS)

We analyzed the HIV Epidemiology Research Study (HERS) to better understand the persistence of infections with specific HPV types. HERS was conducted on 871 HIV seropositive and 439 HIV seronegative women in four US sites. Eligible subjects had HIV without any AIDS defining conditions or were at risk of HIV infection due to intravenous drug use or high risk sexual behavior. Each scheduled visit was six months apart and women were followed for a maximum of 15 visits. Each visit included Pap test screening, gynecological exam and cervicovaginal lavage to collect HPV DNA samples. To be included in the analysis we analyzed results from women that had at least two study visits. Only the first 10 visits from HIV seropositive women were analyzed in this chapter.

We performed analyses on data from the HERS cohort which included 571 women who were HIV seropositive at baseline and had two or more study visits. Of the 571 women, 469 women had negative results at study entry for type 16 and related types and 509 had negative results at study entry for type 18 and related types. We estimated the persistence of HPV type 16 and related types (i.e., HPV types 31, 33, 35, 52, 58) and HPV type 18 and related types (i.e., HPV types 39, 45, 59, 68). Results are presented for “incident” (none of the infections of interest present at baseline) only infections as well as combined “prevalent” (infections present at baseline) and incident infections. Table 5.1 presents the frequency and percentage of high risk human papillomavirus types that were used to estimate persistence.
Several mixed-effect models were used to find a model with the best fit to the data using AIC (Table 5.2). Each model assumed that states 0 and 2 were Markov and the transition probabilities were the same for each type. Models varied on assumptions concerning state 1: Markov and similar among types; Markov and different among types; semi-Markov and similar among types; and semi-Markov and different among types. Also, models that assumed a specific type differed from the other types were fit individually.

For incident type 16 related infections, the best fitting model was Model 9 which assumed state 1 was semi-Markov and transition probabilities leaving state 1 are different for type 52 versus other related types (Figure 5.1); however, when both incident and prevalent type 16 related infections were modeled the best fitting model was Model 2 which assumed state 1 was Markov and transition probabilities leaving state 1 differed among types (Figure 5.2).

For incident type 18 related infections, the best fitting model was Model 3 which assumed state 1 was semi-Markov and the transition probabilities leaving state 1 were the same among types (Figure 5.3). When both incident and prevalent type 18 related infections were modeled the best fitting model was Model 14 which assumed state 1 was semi-Markov and the transition probabilities leaving state 1 are different for type 59 versus other type 18 related types (Figure 5.4).

A likelihood ratio test can be used to test the null hypothesis that there is no random effect, i.e., $H_0: \sigma = 0$. Because under the null hypothesis $\sigma$ is on the boundary of the parameter space, the likelihood ratio test statistic does not have the usual $\chi^2_1$ distribution. Rather, the statistic is a 50-50 mixture of $\chi^2$ distributions with degrees of freedom zero and one. For the HERS data, there was strong evidence of correlation of persistence of different HPV infections within the same person (all p-values $<0.001$). The persistence patterns from mixed effects Model 5 were compared with a fixed effects only model proposed by Mitchell et al. (2011) and an EE proposed by Koshiol et al. (2006a). The estimates are plotted in Figure 5.5.
5.3.2 Placebo arm of two vaccine clinical trials

The placebo arm of two placebo-controlled, double-blind, randomized Phase II/III clinical studies (Protocols 005 and 012) included women age 16-26, who were enrolled and vaccinated without pre-screening for the presence of HPV infection. There were 1040 women from Protocol 005 and 1690 women from Protocol 012 with two or more study visits. Of the 1040 women, 874 of them were HPV negative for all types (i.e., types 6, 11, 16, and 18) at study entry under Protocol 005. Of the 1690 women, 1529 of them were HPV negative for types 16 and 18 at study entry under Protocol 012. Infection status for types 6, 11, 16, and 18 was recorded on Day 1 and Months 7, 12, 18, 24, 30, 36, 42, and 48 under Protocol 005; however, under Protocol 012 only types 16 and 18 were collected and Month 42 was not a scheduled visit. We analyzed the trials separately. To be included in the analysis we analyzed results collected on Day 1 and Months 7, 12, 18, 24, 30, and 36 from women that had at least two study visits and baseline HPV DNA samples for all types with negative results. Table 5.3 presents the frequency and percentage of high risk human papillomavirus types that were used to estimate persistence.

Several mixed-effect models were used to find a model with the best fit to the data using AIC (Table 5.4). Each model assumed that state 0 was Markov and the transition probability leaving state 0 was the same for each type. Models varied on assumptions concerning state 1: Markov and similar among types; Markov and different among types; semi-Markov and similar among types; and semi-Markov and different among types. Also, models that assumed a specific type differed from the other types were fit individually.

For Protocol 005 the best fitting model was Model 4 which assumes state 1 is semi-Markov and transition probabilities leaving state 1 are different among types (Table 5.4). For Protocol 012 the best fitting model was Model 3 which assumes state 1 is semi-Markov and transition
probabilities leaving state 1 are similar among types (Table 5.4). The estimates of HPV persistence from Protocols 005 and 012 are plotted in Figures 5.6 and 5.7.

5.4 Simulation Studies

A simulation study was conducted to assess the bias of the MLE of persistence described above. For the simulation study, persistence is defined as the time between the date of occurrence of the first HPV positive sample and the date of the first occurrence of two consecutive visits recording negative samples. Data sets, each with a sample size of 570 individuals, were randomly generated to be similar to the HERS data. For each individual six types with 10 visits each were generated and used in the analysis. The total number of possible observed time points after study entry was $n_t = 9$. The stochastic process transitioning from state 0 (HPV negative) was Markov and transitioning from state 1 (HPV positive) was Markov. It was assumed that the associated transition probability was the same across HPV types. A similar pattern of missing results used in the simulation studies presented in Chapters 3 and 4 was used.

Simulations were based on fitting the two-state model from Section 5.2.1 to the HERS data for incident type 16 and related HPV infections (i.e., HPV types 31, 33, 35, 52, 58). The parameters were $p_{01} = 0.01$, $\beta_{1j} = 0.65$ for all $j = 1, 2, \ldots, n_J$, $\eta_t = -0.17$ for all $t = 1, 2, \ldots, n_t$ and $\sigma = 1.65$. This results in transition probabilities of $p = (0.01, 0.62)$.

For each simulated data set, the MLEs of the probability of incident type 16 and related HPV type infection persisting at least $t$ years ($t \in \{0.5, 1.0, \ldots, 3.5\}$) were evaluated. The MLEs were computed based on equation 5.4. The average estimates over the 1000 simulations were calculated and the bias is presented in Tables 5.5 and 5.6. The MLEs were approximately unbiased.
Additional simulations were conducted to assess the type I error and power of the likelihood ratio test of the null hypothesis $H_0: \sigma = 0$ described at the end of Section 5.3.1. One thousand simulated data sets were constructed under the null hypothesis $\sigma = 0$ and the empirical type I error of the likelihood ratio test was 0.049. Data sets were also simulated under alternative hypotheses with $\sigma$ equal to values of 0.5, 1.0, 1.5, and 2.0. The empirical power was 0.52, 0.67, 0.78, and 0.97 respectively.

5.5 Discussion

In this chapter we considered mixed effects semi-Markov models to estimate the persistence of high-risk type-specific HPV infections among HIV seropositive women in HERS. These models fully utilize the data since they analyze each type simultaneously. The models include a random effect to account for possible correlation in the duration of different type-specific HPV infections within the same individual. The random effect model requires numerical integration to evaluate the likelihood. Therefore, maximizing the likelihood can be computationally intensive, especially for data sets with many HPV types and models allowing for different persistence distributions between types.
Table 5.1: Frequency and percentage of high risk human papillomavirus types from the 571 HIV-seropositive females in the HIV Epidemiology Research Study with at least two visits

<table>
<thead>
<tr>
<th>HPV Type</th>
<th>Incident Infections</th>
<th>All Infections</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HPV 16 and related</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>49 (12.4)</td>
<td>67 (11.7)</td>
</tr>
<tr>
<td>31</td>
<td>34 (8.6)</td>
<td>54 (9.4)</td>
</tr>
<tr>
<td>33</td>
<td>31 (7.8)</td>
<td>52 (9.1)</td>
</tr>
<tr>
<td>35</td>
<td>34 (8.6)</td>
<td>40 (7.0)</td>
</tr>
<tr>
<td>52</td>
<td>58 (14.7)</td>
<td>77 (13.5)</td>
</tr>
<tr>
<td>58</td>
<td>44 (11.1)</td>
<td>69 (12.1)</td>
</tr>
<tr>
<td><strong>HPV 18 and related</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>48 (12.2)</td>
<td>73 (12.8)</td>
</tr>
<tr>
<td>39</td>
<td>11 (2.8)</td>
<td>14 (2.4)</td>
</tr>
<tr>
<td>45</td>
<td>35 (8.9)</td>
<td>58 (10.1)</td>
</tr>
<tr>
<td>59</td>
<td>14 (3.5)</td>
<td>18 (3.1)</td>
</tr>
<tr>
<td>68</td>
<td>37 (9.4)</td>
<td>50 (8.7)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>395 (100.0)</td>
<td>572 (100.0)</td>
</tr>
</tbody>
</table>
Table 5.2: Model selection for HERS analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Type 16 Related</th>
<th>Type 18 Related</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incid. Infect.*</td>
<td>All Infect.**</td>
</tr>
<tr>
<td>1</td>
<td>2930</td>
<td>4306</td>
</tr>
<tr>
<td>2</td>
<td>2918</td>
<td>4296</td>
</tr>
<tr>
<td>3</td>
<td>2734</td>
<td>4302</td>
</tr>
<tr>
<td>4</td>
<td>2678</td>
<td>4256</td>
</tr>
<tr>
<td>5</td>
<td>2716</td>
<td>4294</td>
</tr>
<tr>
<td>6</td>
<td>2722</td>
<td>4304</td>
</tr>
<tr>
<td>7</td>
<td>2724</td>
<td>4288</td>
</tr>
<tr>
<td>8</td>
<td>2718</td>
<td>4294</td>
</tr>
<tr>
<td>9</td>
<td>2714</td>
<td>4292</td>
</tr>
<tr>
<td>10</td>
<td>2724</td>
<td>4298</td>
</tr>
</tbody>
</table>

* Models assume state 0 is Markov and transitions leaving state 0 are similar among types.  
** Models assume state 0 and 2 are Markov and transitions leaving states 0 and 2 are similar among types.  
Model 1 assumes state 1 Markov and transitions leaving state 1 are similar among types.  
Model 2 assumes state 1 Markov and transitions leaving state 1 are different among types.  
Model 3 assumes state 1 semi-Markov and transitions leaving state 1 are similar among types.  
Model 4 assumes state 1 semi-Markov and transitions leaving state 1 are different among types.  
Model 5 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 16 than the other related types.  
Model 6 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 31 than the other related types.  
Model 7 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 33 than the other related types.  
Model 8 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 35 than the other related types.  
Model 9 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 52 than the other related types.  
Model 10 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 58 than the other related types.  
Model 11 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 18 than the other related types.  
Model 12 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 39 than the other related types.  
Model 13 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 45 than the other related types.  
Model 14 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 59 than the other related types.  
Model 15 assumes state 1 is semi-Markov and transitions leaving state 1 are different for type 68 than the other related types.
Table 5.3: Frequency and percentage of human papillomavirus types from the placebo-arm of two vaccine clinical trials for women that were HPV negative for all types at baseline and had two or more visits

<table>
<thead>
<tr>
<th>HPV Type</th>
<th>No. of Infections (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protocol 005</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>92 (30.0)</td>
</tr>
<tr>
<td>11</td>
<td>23 (7.5)</td>
</tr>
<tr>
<td>16</td>
<td>141 (45.9)</td>
</tr>
<tr>
<td>18</td>
<td>51 (16.6)</td>
</tr>
<tr>
<td>Total</td>
<td>307 (100.0)</td>
</tr>
<tr>
<td>Protocol 012</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>266 (71.7)</td>
</tr>
<tr>
<td>18</td>
<td>105 (28.3)</td>
</tr>
<tr>
<td>Total</td>
<td>371 (100.0)</td>
</tr>
</tbody>
</table>
Table 5.4: Model selection for analysis of Protocols 005 and 012

<table>
<thead>
<tr>
<th>Model*</th>
<th>Protocol 005</th>
<th>Protocol 012</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$-2\ell$</td>
<td>AIC</td>
</tr>
<tr>
<td>1</td>
<td>3870</td>
<td>3876</td>
</tr>
<tr>
<td>2</td>
<td>3826</td>
<td>3838</td>
</tr>
<tr>
<td>3</td>
<td>3468</td>
<td>3484</td>
</tr>
<tr>
<td>4</td>
<td>3404</td>
<td>3456</td>
</tr>
<tr>
<td>5</td>
<td>3446</td>
<td>3474</td>
</tr>
<tr>
<td>6</td>
<td>3440</td>
<td>3468</td>
</tr>
<tr>
<td>7</td>
<td>3430</td>
<td>3458</td>
</tr>
<tr>
<td>8</td>
<td>3462</td>
<td>3490</td>
</tr>
</tbody>
</table>

* Models assume state 0 is Markov and transition probabilities leaving state 0 are similar among types.
  Model 1 assumes state 1 Markov and transition probabilities leaving state 1 are similar among types.
  Model 2 assumes state 1 Markov and transition probabilities leaving state 1 are different among types.
  Model 3 assumes state 1 semi-Markov and transition probabilities leaving state 1 are similar among types.
  Model 4 assumes state 1 semi-Markov and transition probabilities leaving state 1 are different among types.
  Model 5 assumes state 1 semi-Markov and transition probabilities leaving state 1 are different for type 6 versus other types.
  Model 6 assumes state 1 semi-Markov and transition probabilities leaving state 1 are different for type 11 versus other types.
  Model 7 assumes state 1 semi-Markov and transition probabilities leaving state 1 are different for type 16 versus other types.
  Model 8 assumes state 1 semi-Markov and transition probabilities leaving state 1 are different for type 18 versus other types.
Table 5.5: Estimated bias of the probability HPV infection persists at least $t$ years from the simulation study described in section 5.4.

<table>
<thead>
<tr>
<th>$t$ (years)</th>
<th>Truth</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.313</td>
<td>0.007</td>
</tr>
<tr>
<td>1.0</td>
<td>0.103</td>
<td>0.005</td>
</tr>
<tr>
<td>1.5</td>
<td>0.034</td>
<td>0.003</td>
</tr>
<tr>
<td>2.0</td>
<td>0.011</td>
<td>0.001</td>
</tr>
<tr>
<td>2.5</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>3.0</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>3.5</td>
<td>0.0004</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Table 5.6: Estimated bias from the simulation study described in section 5.4.

<table>
<thead>
<tr>
<th>Parameter (years)</th>
<th>Truth</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{01}$</td>
<td>0.01</td>
<td>-0.00007</td>
</tr>
<tr>
<td>$\beta_{1j}$</td>
<td>0.65</td>
<td>0.011</td>
</tr>
<tr>
<td>$\eta_k$</td>
<td>-0.17</td>
<td>0.013</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>1.65</td>
<td>0.015</td>
</tr>
</tbody>
</table>
Figure 5.1: Semi-Markov maximum likelihood estimators (MLEs) of HPV type-16 related persistence from a model assuming transition probabilities for leaving state 1 are semi-Markov and type 52 differs from types 16, 31, 33, 35, and 58 for women in HERS cohort who were HIV positive and HPV negative at study entry.
Figure 5.2: Markov maximum likelihood estimators (MLEs) of HPV type-16 related persistence from a model assuming transition probabilities for leaving state 1 are Markov and different among types 16, 31, 33, 35, 52 and 58 for women in HERS cohort who were HIV positive at study entry.
Figure 5.3: Semi-Markov maximum likelihood estimators (MLEs) of HPV type-18 related persistence from a model assuming transition probabilities for leaving state 1 are semi-Markov and the same for types 18, 39, 45, 59 and 68 for women in HERS cohort who were HIV positive and HPV negative at study entry.
Figure 5.4: Semi-Markov maximum likelihood estimators (MLEs) of HPV type-18 related persistence from a model assuming transition probabilities for leaving state 1 are semi-Markov and type 59 differs from types 18, 39, 45 and 68 for women in HERS cohort who were HIV positive at study entry.
Figure 5.5: Empirical estimator and semi-Markov maximum likelihood estimators (MLEs) with and without random effects of HPV type-16 persistence among women in HERS cohort who were HIV positive and HPV type 16 negative at study entry. Mixed effects model 5 was used.
Figure 5.6: Semi-Markov maximum likelihood estimators (MLEs) of HPV persistence from a model assuming transition probabilities leaving state 1 are semi-Markov and types 6, 11, 16 and 18 are different for women in the placebo arm of a vaccine clinical trial (Protocol 005) who were HPV negative for types 6, 11, 16 and 18 at study entry.
Figure 5.7: Semi-Markov maximum likelihood estimators (MLEs) of HPV persistence from a model assuming transition probabilities leaving state 1 are semi-Markov and types 16 and 18 are the same for women in the placebo arm of a vaccine clinical trial (Protocol 012) who were HPV negative for both types at study entry.
Chapter 6

CONCLUSION

Obtaining accurate and precise estimates of HPV persistence has important public health implications. The future formulation of policies for inclusion of HPV testing in cervical cancer screening and prevention depends on clearly defining and estimating persistence. Developing consensus definitions and estimators can usefully inform clinical practice for future cervical cancer screening programs. This thesis sought to develop semi-Markov discrete-time multi-state models in order to accurately estimate HPV persistence under minimal assumptions.

In Chapter 3, we considered survival analysis methods (i.e., EEs) commonly used in the HPV literature as well as MLEs of type-specific HPV persistence using panel data on individuals that do not have the type-specific infection at study enrollment. The proposed MLE used a semi-Markov two-state discrete-time model for incident infections only. These estimators were compared in simulation studies as well as applied to the HERS data. Advantages of the MLE of HPV type-specific persistence include that it is nonparametric, does not require guarantee times, does not require the Markov assumption, is approximately unbiased (if model is specified correctly and MAR assumption holds), and flexible with respect to estimands. On the other hand our results show that commonly used EEs are biased, and that bias can be quite large in practice. The most commonly used formal definition of HPV persistence is HPV positivity at a minimum of two consecutive follow-up visits with a median interval of six months apart (Rostich et al. (2011)). The requirement of two consecutive negative visits after positivity is also common for EEs and was used to define persistence in Chapter 3. It
is the requirement of two consecutive negative results that increases the bias of the EE. The Markov assumption was tested for the HERS data and it was concluded that the semi-Markov assumption was more appropriate. Even though the Markov assumption is popular and simpler, it was not appropriate in modeling the HERS data.

In Chapter 4, we proposed an extension of the semi-Markov MLE (i.e., SSPLA) from Chapter 3 in order to include prevalent infections and account for length bias sampling. The SSPLA estimator was compared to the IO and NSME estimators in simulation studies as well as applied to the HERS data. We concluded that the NSME MLEs were biased. The SSPLA and IO estimators were nearly unbiased and similar when the prevalence of infection was fairly small. Simulation studies showed that the efficiency of the SSPLA estimator relative to the IO estimator increased as baseline prevalence increased. While it is always appropriate to use SSPLA, the IO method is suitable to use when prevalence is low since one can obtain similar estimates to SSPLA without the added complexity of using nuisance parameters.

In Chapter 3 type-specific incident infections were analyzed separately and in Chapter 4 multiple incident and prevalent infections were analyzed as a composite endpoint combining any HPV infection. However, in Chapter 5 the semi-Markov MLE was further extended to account for clustering when studying the association between HPV type and persistence of HPV infections. Simulation studies showed that the MLE is nearly unbiased. The results from the HERS data varied. We found that persistence of incident type 16 infections differed from the persistence of other type 16 related incident infections and the Markov assumption was not appropriate; however, when analyzing all type 16 and related infections, the persistence was the same across types and the Markov assumption was valid.

Discrete time semi-Markov models can be used to analyze incident and prevalent HPV infections by type individually or as multiple types simultaneously. One future research direction is to apply these methods to other biomedical panel data in settings where infections reoccur
or reactivate. The models proposed here did not assess the relationship of baseline or demo-
graphic characteristics with persistence. Another future research direction is to add covariates
of interest into the models to assess these relationships.
Bibliography


