Topics in Current Status Data

by

Karen Michelle McKeown

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor in Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nicholas P. Jewell, Chair
Professor Alan E. Hubbard
Professor John M. Colford

Fall 2011

Topics in Current Status Data

Abstract

Topics in Current Status Data

by

Karen Michelle McKeown

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Nicholas P. Jewell, Chair

This dissertation considers topics in current status data, a type of survival data where the only available information on the survival time is whether or not the event time has occurred before the examination time. We introduce the concept of current status data and give some motivating examples to highlight some of the many areas in which this type of data naturally occur in practice. We discuss some of the well known and widely used methods for analyzing current status data, along with some of the more recent developments in the area, and provide appropriate references to these previously examined methods. Within this dissertation, we add to the existing literature in the area by developing ideas not previously addressed from a current status data perspective.

We describe a simple method for nonparametric estimation of a distribution function based on current status data where observations of current status information are subject to (known) misclassification. Nonparametric maximum likelihood techniques are obtained through the use of a straightforward set of adjustments to the familiar pool-adjacent violators algorithm, which is generally used when misclassification is assumed absent. The methods are extended to allow for misclassification rates that vary over time, particularly when misclassification is most likely to occur close to the time of the true failure event. Using the ideas of binary generalized linear models with outcomes subject to misclassification we consider regression models for the underlying survival time. The ideas are motivated by and applied to an example on human papillomavirus (HPV) infection status amongst women examined in San Francisco. Additional applications on breastfeeding behaviors and menopausal status are also presented. As an extension we consider group testing with current status data in the presence of misclassification. Group testing combines samples, such as blood or urine, from a number of individuals and tests the group sample for the presence of the disease of interest instead of testing each individual sample. We examine whether group testing can be used to not only reduce the costs incurred with testing a large number of individuals but also improve the efficiency in estimating the underlying distribution function. We also seek to determine the optimal group size for nonparametric estimation of a distribution function, under various group testing scenarios. Regression models for the group testing approach are briefly considered.

We also describe current status data from the perspective of counting processes. We examine the relationship between current status data and simple counting processes. Specifically we consider the multistate model defined by two survival times of interest where one only observes whether or not each of the individual survival times exceed a common observed monitoring time. We are interested in estimation of the distribution function of time to the first event and whether current status information on the subsequent event can be used to improve this estimate. For both single and multiple monitoring time scenarios, in the fully nonparametric setting, one cannot improve the naïve estimator, using information on the first event only, when estimating smooth functionals of the distribution of time to the first event (van der Laan and Jewell, 2003). We therefore examine improving this naïve estimator when parametric assumptions about the waiting time between the two events are made. For situations where this waiting time is modifiable by design, we also determine the optimal length of the waiting time for estimation of the cumulative hazard of the distribution of time to the first event in the recent past. The ideas are motivated by and applied to an example on simultaneous accurate and diluted HIV test data.

This dissertation is dedicated to my parents, Aidan and Bernadette McKeown, without their love and support none of this would be possible.

Contents

List of Figures

List of Tables

Acknowledgements

There are many important people I would like to thank for the support and encouragement they have given me over the years, some of whom may not realize the impact they have had on completing this dissertation.

This dissertation would not have been possible without the help of my adviser, Nicholas P. Jewell, with whom it has been an honor to work with. I am thankful not only for the time and advice he has given me regarding this dissertation, but for his genuine support throughout my years at U.C. Berkeley. The benefits of his guidance will remain with me long after I graduate. I will be forever grateful for the opportunities he has made possible for me.

I wish to thank the other members of my dissertation committee, Alan Hubbard and Jack Colford, for reviewing this dissertation and for the new ideas, extensions, and helpful suggestions they have given.

I would like to acknowledge the National Institute of Allergy and Infectious Diseases for their support of my research through grant #R01-ES015493.

I would like to express my appreciation to Sharon Norris and Burke Bundy for their hard work and dedication. Sharon's quick response to all questions and concerns, along with Burke's patience in solving all my computer problems have given me the ability to focus on writing this dissertation without additional distractions, for which I am extremely grateful.

My parents, Aidan and Bernadette McKeown, deserve special recognition for their encouragement and support from the very beginning. They have given me every possible opportunity throughout my life and the courage to follow this dream, for this I am eternally thankful. I would also like to thank my family and friends in Ireland for being with me every step of the way, despite being so far away. A special thanks must also be given to my grandmother, Marion McKeown, for her endless prayers, candles and phone calls which have helped me along this journey.

I would like to recognize the support I have received from all the lifelong friends I have made while at U.C. Berkeley. They filled my days with laughter and gave me the motivation and encouragement needed to complete my research.

I cannot thank my fiance, Shourya Basu, enough for his contribution to this dissertation. His unwavering strength and reassurance have allowed me to stay steady during challenging times. I cannot imagine completing this dissertation without his love, optimism, and endless help and support. I also wish to thank his family for giving me a home away from home.

Overview

Analyzing survival data is complicated by issues associated with incomplete data, some of which arise due to various forms of censoring. Current status data, sometimes referred to as interval censoring case I (Groeneboom and Wellner, 1992), is an extreme form of censoring. Current status data provides information on the survival status of individuals at various times rather than standard observation of failure times, or possibly right censored failure times. As stated by Jewell and van der Laan (1995), interest in current status data has arisen because of a number of interesting applications, the association between current status data and related problems in deconvolution, and the nature of the non-standard results associated with the non-parametric maximum likelihood estimate (NPMLE). Chapter 1 summarizes and extends the work of Jewell and van der Laan (2004b) to give a review of the existing methods used with current status data, and to highlight recent developments in the area. Aspects such as regression, asymptotic properties, competing risks, cure models, and various current status data sampling schemes are addressed. Subsequent chapters of this dissertation aim to add to the existing literature on the analysis of current status data by addressing issues not previously examined.

Chapter 2 considers misclassification of current status data where the current status outcomes are subject to misclassification. We extend the known nonparametric maximum likelihood estimator of the distribution function underlying current status data when there is no misclassification (Ayer et al., 1955) to allow for time-independent misclassification, with known misclassification rates. We consider the potential impact of misclassification rates that vary over time, in particular when misclassification only occurs in a known time window surrounding the underlying failure event. Using the ideas for binary generalized linear models with outcomes subject to misclassification (Neuhaus, 1999), we also consider regression models for current status data subject to misclassification. Chapter 3 further extends Chapter 2 where we consider misclassification when the individual samples are combined for group testing, an approach which is used to reduce the amount of testing needed without screening fewer individuals. We determine the optimal group size, assuming known misclassification rates, for estimating disease prevalence. Chapter 4 considers current status observation of simple counting processes. We focus primarily on current status observation of a three state counting process, defined by the occurrence of two events. We are interested in estimating the time until the first event, and specifically whether current status information on a subsequent event can be used to improve this estimate. Chapter 5 gives an overall summary of the motivation behind this dissertation.

Chapter 1

Current Status Data Review

1.1 Background

When analyzing data, observations may be subject to censoring. Censoring occurs when the value of the observation is only partially known. In survival analysis, right censored data is commonly observed. Right censored data arises when the value of an observation is greater than a certain point, but the amount by which it is greater is unknown. This situation occurs naturally in medical and public health examples where individuals are examined for the occurrence of an event of interest over a pre-specified period. At the end of this time period some of the individuals may not have experienced the event. It is therefore known that the event time for these individuals is greater than the length of the time interval, but the exact amount greater is not known. Interval censored data is another common data structure used in estimation of survival data. Interval censored data arises when a failure time cannot be directly observed, instead it can only be determined to lie within an interval. This type of data is usually obtained through the use of repeated examinations on each individual. The exact timing of the event is not observed but the event is determined to lie between two successive examination times. See Huang and Wellner (1997) or Zhang and Sun (2010) for an overview of interval censored data. In many situations, due to time constraints, costs, and other difficulties, a more extreme form of censored survival data may be preferred, namely current status data, or Type I interval censored data (Groeneboom and Wellner, 1992). With current status data the only available information on the failure time is whether or not it has occurred before the examination time. With current status data we never observe the exact failure time for any individual, therefore differing greatly from right censored data.

In the next subsection we introduce some motivating examples which highlight some of the many areas in which current status data occur naturally in practice. Specific examples are given in Chapters 2, 3 and 4, however, many other examples would be equally applicable to the statistical methods of these chapters, as seen in the additional data applications of Appendix A. Although there remain concepts applicable to current status data not addressed here, this chapter gives an overview of some of the key concepts applied to current status data and gives the reader information and references on where to obtain a more detailed description of the methods presented. We hope that this chapter will serve as an informative starting point for those interested in exploring the analysis of current status data in more detail.

1.1.1 Motivating Examples

There are many advantages to using current status data, the most obvious of which is that current status data is less costly and time consuming than obtaining longitudinal data. Earliest work with current status data was motivated by applications in demography, where the distribution of age at weaning is commonly studied under various settings (Diamond et al., 1986, Diamond and McDonald, 1991, Grummer-Strawn, 1993). In such an application, inaccuracy and bias surrounding the retrospective recall of the exact timing of an event, such as the time of weaning, has led to use of current status data as the preferred form of data collection for understanding the distribution of time to a specific event, in this case, the age at weaning. The bias and inaccuracy introduced with self-reporting is addressed in more detail throughout Chapter 2.

Early work with current status data has also been applied to carcinogenicity studies (Gart et al., 1986). Since this initial use, current status data has been continually applied to carcinogenicity testing (Lin et al., 1998, Zhang et al., 2005, Tong et al., 2007). Carcinogenicity testing is used when there is an occult tumor under investigation and the distribution of time from exposure to a potential carcinogen until the development of a tumor is of interest. The presence or absence of the occult tumor is determined through animal sacrifice, thus providing current status information at the time of sacrifice. Current status data has also been applied to other areas of epidemiology (Becker, 1989).

Partner studies of Human Immunodeficiency Virus (HIV) transmission have also received considerable attention from a current status data framework (Jewell and Shiboski, 1990, Shiboski and Jewell, 1992, Shiboski, 1998b). A straightforward HIV partner study collects HIV infection data on both partners in a long-term sexual relationship. It is assumed that each partnership consists of a primary infected individual, the index case, who became infected through an external source. The time from infection of this index case to infection of a susceptible partner is then of interest. A necessary assumption is that the susceptible partner has no other means of infection other than contact with the index case.

Age-incidence estimation for a non-fatal human disease can also be obtained through the use of current status data. This approach is useful when the exact incidence of the disease of interest is not known but where disease prevalence can be determined through the use of an accurate diagnostic test. Keiding (1991) estimates the age at incidence of Hepatitis A based on a cross sectional sample of a given population examined by an accurate diagnostic test. The use of current status techniques in this setting is further examined by Keiding et al. (1996). For a rare disease, a case-control sampling scheme, described in Section 1.4.1, should be used for this approach to age-incidence estimation.

Other examples of current status data are used throughout this dissertation to highlight specific statistical concepts and ideas. Although an understanding of the biology and history of the diseases of interest are not essential for understanding the statistical issues and solutions presented, a basic knowledge helps the reader appreciate the problem being addressed. To give the reader a more general understanding, a brief overview of the diseases used throughout this dissertation are given in the appendices. These appendices are not intended to give the reader a complete and comprehensive knowledge of the disease, instead a more general understanding of how the disease affects individuals, the signs, symptoms, available vaccines, and methods of prevention are identified. They also seek to highlight the reasons for and importance of studying such illnesses. Specifically, Appendix B and Appendix C describe the Human Papillomavirus (HPV) and the Human Immunodeficiency Virus (HIV).

1.1.2 Notation and Likelihood

For simple current status data, let T be the survival time random variable of interest with the corresponding distribution function F, and associated survival function $S = 1 - F$. Let the monitoring time, the time at which an individual is examined, be denoted by the random variable C. With current status data we do not observe the exact failure times, instead we observe information on the survival status of an individual at a specific point in time, C. Observation of the survival time random variable, T, is therefore restricted to knowledge of whether or not T exceeds the monitoring time, C. When C is random, as is often the case, the data can be represented by n observations from the joint distribution of (T, C) where n i.i.d. observations are collected on the random variable (Y, C) with $Y = I(T \leq C)$. In general, C is assumed to be independent of T. Section 1.8 briefly considers the case where C may depend on the underlying survival time T, however, throughout this dissertation we assume C is independent of T.

When the monitoring time C is random, C is assumed to follow a distribution function G. For convenience we generally assume C is random. However, most techniques, including those used in this dissertation, are based on the conditional distribution of T, given C, and can therefore also be applied to the case of fixed, non-random C. In the random case, it is often assumed that the data arise from a simple random sample from the joint distribution of T and C. Much of the focus of this dissertation is based on this i.i.d. sampling of (Y, C). Section 1.4 outlines some alternative sampling schemes which are also equally applicable to the current status data structure. Appropriate references are contained within where a more detailed description of each form of current status data can be obtained.

A great deal of interest is given to non-parametric estimation, and inference, of F. With simple current status data, since the binary random variable Y takes the value 1 when $T \leq C$ and 0 when $T > C$, estimation of F can be viewed in terms of estimation of the conditional expectation of Y, for all c, with a monotonicity constraint imposed on the regression function. This can be seen mathematically by $E(Y|C = c) = P(T \leq C|C = c) = F(c)$. If an i.i.d. sample of n individuals is obtained, the likelihood of this data is given by

$$L = \prod_{i=1}^{n} F(c_i)^{y_i}(1 - F(c_i))^{1-y_i} dG(c_i),$$

where c_i and y_i are the observed values of the monitoring time and current status outcome, for individual i, respectively. For the random monitoring time scenario, with C independent of T, estimation of F can then be based on the conditional likelihood of Y, given C. This conditional likelihood is also applicable when C is fixed and non-random, again assuming C is independent of T. The conditional likelihood is given by

$$CL = \prod_{i=1}^{n} F(c_i)^{y_i}(1 - F(c_i))^{1-y_i}. \tag{1.1}$$

The following section considers estimation of the non-parametric maximum likelihood estimator of F based on this likelihood, Equation (1.1). Considerable attention is given to estimation of a survival function (or distribution function) in subsequent chapters, under various scenarios. Section 1.3 considers regression models with current status data. Although this i.i.d. sampling scheme is commonly used for current status data, Section 1.4 identifies other sampling schemes which can benefit from current status data techniques. Other interesting areas of current status data are also considered later in this chapter, including cure models (Section 1.6), competing risks current status data (Section 1.7), and misclassification of current status data (Section 1.9).

1.2 Estimation of a Survival Function

Attention has been given to estimation of a survival function, S, or the corresponding distribution function, F, based on current status data, for both the simple current status data structure given in Equation (1.1), and more complex current status data structures. Estimation of a variety of functionals of F has also been considered. In the single-sample setting, the likelihood of Equation (1.1) can be maximized over the space of all distribution functions to obtain the nonparametric maximum likelihood estimate (NPMLE) of F. The pool-adjacent violators (PAV) algorithm of Ayer et al. (1955) is a widely used algorithm to obtain this nonparametric maximum likelihood estimator, $\widehat{F}$, more information on which is given below. Barlow et al. (1972) and Groeneboom and Wellner (1992) discuss the connection between this estimator and convex minorants.

It is important to realize that with current status data, the data obtained can only identify the value of the distribution function F at the observed monitoring times, $c_1, c_2, ..., c_n$. The corresponding estimator, $\widehat{F}$, is a step function which can jump at each observed monitoring time, or a subset of monitoring times, but cannot jump at any other time. The PAV algorithm is a simple iterative algorithm used for solving monotonic regression problems. Although this relatively straightforward algorithm has been widely used and explained, we rely heavily on this algorithm throughout this dissertation and therefore present a summary of the steps involved in this algorithm in Figure 1.1. See de Leeuw et al. (2009) for a more complete overview of monotonic regression, and how to implement such an algorithm using available statistical software.

Step 1: Order the n observations (C_i, Y_i) by increasing monitoring times to obtain $(C_{(i)}, Y_{(i)})$, where $C_{(1)} \leq C_{(2)} \leq ... \leq C_{(n)}$.
Let $i = 1$.

Step 2: If $Y_{(i)} > Y_{(i+1)}$, thus violating the monotonicity constraint, perform adjacent pooling and replace $Y_{(i)}$ and $Y_{(i+1)}$ by the average;

$$Y^*_{(i)} = Y^*_{(i+1)} = (Y_{(i)} + Y_{(i+1)})/2.$$

Step 3: If $Y_{(i-1)} \leq Y^*_{(i)}$, let $i = i + 1$ and repeat *Step 2*.
If $Y_{(i-1)} > Y^*_{(i)}$, pool $\{Y_{(i-1)}, Y_{(i)}, Y_{(i+1)}\}$ into one average. Continue pooling until $Y_{(k)} \leq Y^*_{(i)}$, for all $k < i$. Then, let $i = i + 1$ and repeat *Step 2*.

Figure 1.1: Illustration of the key steps involved in implementing the Pool-Adjacent Violators (PAV) Algorithm of Ayer et al. (1955).

It is interesting to note that if the data does not violate the monotonicity constraint at any of the monitoring times, the PAV algorithm will simply reproduce the original data as its output. The presence of outliers in the data will cause the PAV algorithm to produce a step function with long, flat levels. From Step 1 of Figure 1.1 we can see that, through the initialization of $i = 1$, the algorithm begins with the first ordered observation to obtain an increasing fit of the NPMLE. A different resulting distribution will be produced if the algorithm begins with the final ordered observation and proceeds in reverse to produce a decreasing fit from this final observation. It is therefore important to note that throughout this dissertation, when we use the PAV algorithm, it is assumed we begin with the first ordered observation, as described in Figure 1.1.

For a hypothetical data set, Figure 1.2 shows how the PAV algorithm can be used to obtain $\widehat{F}$, the NPMLE of F. Here we assume there are six distinct ordered monitoring times, $C_{(1)}, C_{(2)}, ..., C_{(6)}$. The initial observed values of $F(C_i)$ are simply the proportion of individuals observed at time C_i with a positive outcome. Figure 1.2(a) shows the true distribution function F as a solid curved line, with the initial observed values of F represented by a solid circle at each of the monitoring times. Figure 1.2(b) adds the pooled estimates to the previous panel, which are represented by open circles. Finally, Figure 1.2(c) shows $\widehat{F}$ as a step function. Again, it can be seen that all jumps occur at a monitoring time.

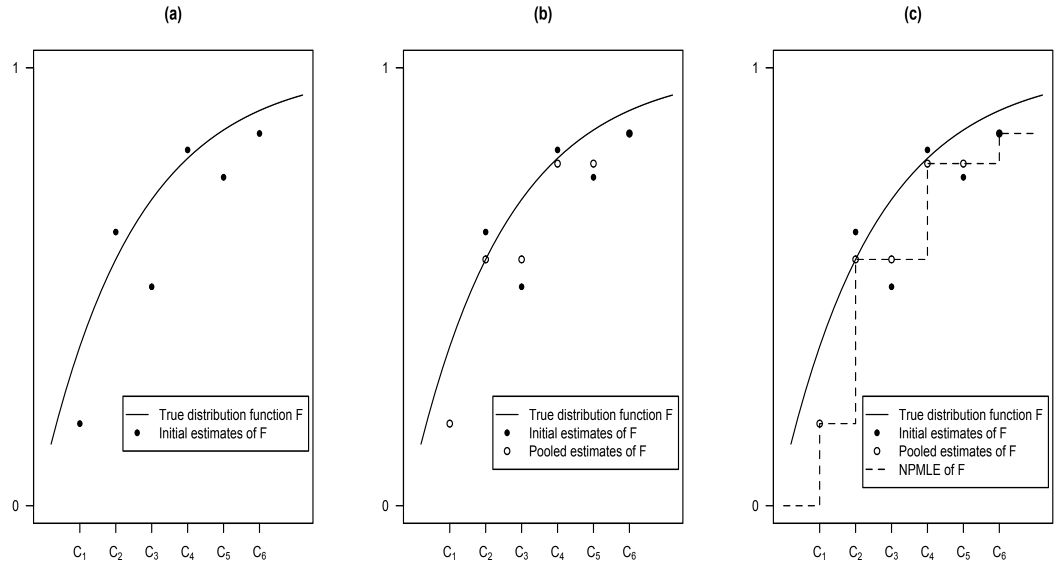

Figure 1.2: Illustration of the use of the Pool-Adjacent Violators (PAV) Algorithm for estimation of the NPMLE of F, using hypothetical data.

Although this straightforward use of the PAV algorithm is the most commonly used approach for nonparametric estimation of a survival function with this type of data, variations of this method also exist, including a weighted version. Burr and Gomatam (2002) study non-parametric estimation of the conditional distribution function, $F(Y|X = x)$, when there is current status information on the outcome variable and a single continuous-valued covariate X. An estimate of this conditional distribution function is based on a locally weighted version of the NPMLE for current status data in the absence of covariates. The asymptotic distribution of this local NPMLE of the conditional distribution function at a single point is also addressed.

1.2.1 Asymptotic Properties

There is by now a growing literature on the non-standard asymptotic properties of the standard NPMLE of F for this simple current status data setting. With simple current status data, the NPMLE of F converges to F at a slower rate of convergence, $n^{1/3}$ as opposed to the familiar $n^{1/2}$ rate of the empirical distribution function or the Kaplan-Meier estimator. The limiting distribution of $\widehat{F}$ is not Gaussian. Groeneboom and Wellner (1992) describe this limit distribution as a complex distribution associated with two-sided Brownian motion. It is therefore not appropriate to focus on the (asymptotic) variance of the non-parametric maximum likelihood estimate of F based on any form of current status data as a step towards confidence interval construction. For pointwise confidence intervals for F, various approaches have been developed for standard current status data (Banerjee and Wellner, 2005). Banerjee and Wellner (2005) propose and discuss the properties of likelihood ratio based confidence intervals. These intervals are compared to the quantiles of the limiting distribution of $\widehat{F}$, estimated using various methods, including parametric fitting, ad-hoc non-parametric procedures and subsampling techniques. In general, the standard bootstrap yields inconsistent estimates of pointwise confidence intervals whether data is sampled with replacement from the original data or generated from the nonparametric maximum likelihood estimator (Sen et al., 2010). As a modification, a smoothed version of the bootstrap is appropriate, as is the m out of n bootstrap (Politis et al., 1999). Further details on the m out of n bootstrap are given in Section 2.2.1, with the ideas applied to a dataset on Human Papillomavirus (HPV) infection in Section 2.2.2.

As stated in Jewell and van der Laan (2004b), despite the unusual and slow rate of convergence, Huang and Wellner (1995) show that estimates of smooth functionals of F, based on $\widehat{F}$, converge at rate $n^{1/2}$ and are asymptotically efficient at many data generating distributions. These authors also supply the influence curve for such smooth functional estimators, thereby facilitating straightforward calculations for (asymptotic) confidence intervals.

1.3 Regression Models

In many applications, information on a (potentially multi-dimensional) covariate X is often also available. The n i.i.d. observations can then be represented by the triplets (Y, C, X), again where Y is the observed current status outcome. Considerable attention has been given to estimation of regression coefficients from a variety of standard models, in both the parametric, semi-parametric and nonparametric setting (Shen, 2011, Shiboski, 1998a, Honda, 2004). Much of the literature on current status data has exploited the correspondence between standard regression models for the underlying failure time and generalized linear models for the observed current status outcome. Doksum and Gasko (1990) initially considered the associated between survival and binary regression models in the context of censored survival data. This approach has since been extended to link the failure time T and the current status outcome Y to the covariate vector X. This approach allows estimates of the parameters in the regression model for the observed Y to be interpreted in terms of the parameters in the regression model for the unobserved survival time T. Jewell and van der Laan (2004b) give examples describing this correspondence, giving a comprehensive overview of current status data in the regression setting. Assuming the survival times follow a proportional hazards model, the current status random variable Y is related to X through a generalized linear model for Y with complimentary log-log link and offset given by an arbitrary increasing function of the observed 'covariate' C. In Section 2.2.5 we consider regression analysis when the current status outcome is subject to misclassification, where the link between generalized linear models and regression with current status data is clearly outlined, both in the presence and absence of misclassification.

If the baseline survival function S_0 is assumed to follow a particular parametric form, the corresponding binary regression model will often simplify to a familiar generalized linear model, so that standard software can be used to estimate both S_0 and the regression parameters. In the application of Section 2.2.5 it is shown that by assuming a Weibull regression model for T, the generalized linear model for Y in X and C involves the complementary log-log link function. When the survival times follow a proportional odds regression model, Y is associated with X via the logit link function. Various regression models have been studied for current status data by other authors, including proportional hazards (Huang, 1996), proportional odds (Rossini and Tsiatis, 1996), additive hazards (Lin et al., 1998, Martinussen and Scheike, 2002), linear transformation (Sun and Sun, 2005), additive transformation (Cheng and Wang, 2011), accelerated failure time (Tian and Cai, 2006), and cure models (Ma, 2009).

Shiboski (1998a) shows that if S_0 is left arbitrary, semi-parametric methods can be used to examine inference on the vector of regression coefficients, λ, treating S_0 as a nuisance parameter. As dependance between C and the covariates X can introduce some bias in estimation of λ, Shiboski (1998a) also describes some simulations that compare the relative performance of coefficient estimates based on parametric or nonparametric assumptions on S_0. Xue et al. (2004) consider a semi-parametric regression model that consists of parametric and nonparametric regression components. A sieve maximum likelihood estimator is proposed to estimate the regression parameter, allowing exploration of the non-linear relationship between a certain covariate and the response function. Asymptotic properties of the proposed estimator are discussed within.

Honda (2004) considers nonparametric regression with current status data where no parametric assumptions are imposed on the distributions of C or X, or the random error, ϵ. It is however assumed that (C, X) and ϵ are independent. A nonparametric estimator of the regression function is obtained by modifying the maximum rank correlation estimator of Han (1987), originally proposed for linear models. The asymptotic bias and asymptotic distribution of the estimator are discussed therein.

1.4 Sampling Schemes

Although an i.i.d. sample of n observations of (Y, C) is a common sampling scheme, there are several applications in which alternative sampling schemes may be preferred. The following subsections introduce different sampling schemes which have received attention from a current status data perspective, including case-control sampling, doubly censored current status data, clustered current status data, and bivariate current status data.

1.4.1 Case-control Current Status Data

As with interval censored data with short intervals, if the failures of interest are infrequent (which occurs when dealing with a rare disease) a random sample of the population may result in very few current status observations where the event of interest has occurred before the monitoring time. When this occurs a case-control sampling scheme is often preferred where separate samples of individuals for whom the event has already occurred (cases) and those for whom the event has not occurred (controls) are obtained. For more information on standard case-control designs see Aschengrau and Seage (2003). With case-control current status data, an identifiability problem arises in nonparametric estimation of F, even when the support of T is contained within the support of C. Jewell and van der Laan (2004a) consider case-control current status data and show that with such data one can only identify the odds function associated with F up to a constant. In an assumed parametric family this may be enough to identify F, but this will not be the case nonparametrically. It is also shown that the NPMLE based on case-control current status data supplemented by knowledge of the number of individuals for whom $T \leq C$ and $T > C$ can be obtained by weighting observations inversely proportional to their probability of selection, and using standard methods for simple current status data on this weighted data.

1.4.2 Case-cohort Current Status Data

A useful modification of the case-control study design is the case-cohort design (Sato, 1992, Miettinen, 1982). As with case-control designs, a separate sample of those for whom the event has already occurred (cases) is obtained. However, with case-cohort studies, the 'controls' now consist of a sample of individuals from the population, or cohort, of initially disease-free individuals. Therefore, in this setting, an individual could be a member of both samples, with the control sample containing some cases. Ma (2007) consider case-cohort sampled current status data under the assumption of an additive risk model. Through the use of estimating equations, estimation of the regression parameters are considered. Asymptotic properties and inference are also considered. Li and Nan (2011) consider semi-parametric regression models for case-cohort current status data, specifically relative risk regression.

This work is approached from a more general perspective of two-phase sampling designs, of which the case-cohort design is a specific example.

1.4.3 Doubly Censored Current Status Data

If interest is on the estimation of the distribution of time between two consecutive events where the initial event time is known, yet the only available information on the subsequent event is whether or not it is greater than an observed point in time, the data is considered singly censored (or simple) current status data. DeGruttola and Lagakos (1989) describe the general case of doubly censored survival data, a special case of which becomes doubly censored current status data (Rabinowitz and Jewell, 1996), where the initial event time is also unknown. Under this data structure the survival time random variable T measures the length of time between two successive events, I and J, measured in chronological time. If individuals are monitored at the single chronological time C, the observed data can then be described as (Y, C) where $Y = 1$ if $J \leq C$ and $Y = 0$ if $I \leq C < J$, therefore producing a form of current status data observing whether either the initiating or subsequent event have occurred before the monitoring time. This definition of the indicator variable Y is useful as if $J \leq C$, this implies $T \leq C - I$. Similarly, if $I \leq C < J$, then $C - I < T$. This type of data structure is useful in applications to data on HIV partner studies (as described in Section 1.1.1) where T is defined as the time between HIV infection of an individual (the index case) and transmission of HIV to their sexual partner.

In examining doubly censored current status data, Jewell et al. (1994) assume the conditional distribution of I, given $I \leq C$, is known and give the conditional likelihood for n observations, from which parametric estimation of F can again be obtained based on standard methods. They also approach nonparametric maximum likelihood estimation of F by viewing the model as a nonparametric mixture estimation problem. This work is extended by other authors to further examine estimation of this NPMLE with doubly censored current status data, and the corresponding asymptotics (van der Laan et al., 1997a, van der Laan and Jewell, 2001). Rabinowitz and Jewell (1996) also examine the use of doubly censored current status data to estimate the regression coefficient in an accelerated failure time model for the length of time between two successive events. van der Laan et al. (1997a) give an asymptotically efficient method for carrying out regression analysis for T. The methods of which are explained in detail for simple linear regression and later generalized for linear and non-linear regression.

1.4.4 Clustered Current Status Data

Clustered data are characterized as data that can be classified into a number of distinct groups or clusters within a particular study (Galbraith et al., 2010). Clustered data arises from many applications, including studies in ophthalmology, family-based genetics, community interventions and biomedical data analysis. An important feature of clustered data is that outcomes from the same cluster are likely to be positively correlated. This within-cluster correlation therefore needs to be taken into consideration when analyzing such survival data. The ignorance of such correlation can bias the statistical inference (Ying and Liu, 2006). The modeling issues that arise during the analysis of clustered failure time data differ slightly from those with the analysis of multivariate failure time data. The dimension of the cluster-level response vector is often much larger than in the multivariate case since it is defined by cluster sizes, which can potentially vary. Clustered event data have been widely examined for right censored data. Current status data can also be clustered, where each individual is observed only once. A cluster may represent a group, such as a family, or may consist of a single subject, for example, when examining the time until blindness in the right and left eyes of a single individual (Wen et al., 2011).

With clustered current status data T_{ij}, C_{ij}, and X_{ij} denote the event time, the monitoring time, and the multi-dimensional covariate vector of the jth member in the ith cluster. Assume there are r clusters with $m_i \geq 2$ members in the ith cluster. The observed data can then be written as $\{(C_{ij}, Y_{ij}, X_{ij}) | i = 1, ..., r, j = 1, ..., m_i\}$, where $Y_{ij} = I(T_{ij} \leq C_{ij})$ is a binary variable, indicating whether the event has occurred before the examination time for the jth member in the ith cluster. Wen and Chen (2011) consider nonparametric maximum likelihood analysis of clustered current status data, assuming a gamma-frailty Cox model. The gamma-frailty model is also considered by Hens et al. (2009) for clustered current status data, where details are given on both the shared frailty model and the correlated frailty model. For more information on frailty models see Vaupel et al. (1979). Cook and Tolusso (2009) consider second-order estimating equations for the analysis of clustered failure time data under a current status observation scheme, where the proportional hazards formulation is assumed to examine covariate effects.

1.4.5 Bivariate Current Status Data

Suppose instead of being interested in a single failure time, there exists two failure times of interest on which only current status information is available for both times. In this scenario with two random survival variables of interest, T_1 and T_2, Wang and Ding (2000) define bivariate current status data as observing whether or not T_j, for $j = 1, 2$, exceeds the random monitoring time C, for each individual. The n i.i.d. observations could therefore be written as (Y_1, Y_2, C), where $Y_1 = I(T_1 \leq C)$ and $Y_2 = I(T_2 \leq C)$. For a thorough description of bivariate current status data see Jewell et al. (2005), where it is shown that for bivariate current status data with univariate monitoring times, the identifiable part of the joint distribution is three univariate cumulative distribution functions, namely the two marginal distributions and the bivariate cumulative distribution function evaluated on the diagonal. The complete bivariate distribution is not identifiable. Some of the simple current status data examples introduced in Section 1.1.1 can be extended by incorporating additional information, resulting in the bivariate current status data structure. For example, with carcinogenicity testing, animals could be tested for the presence of non-lethal tumors at two different sites. With the HIV partner studies example, T_1 could define the time until HIV infection of the partner, with T_2 defining the time until the index case develops AIDS.

Many of the problems of interest considered for simple current status data are also of interest for bivariate current status data. With the additional information available in bivariate current status data comes additional problems of interest. In the bivariate current status data setting, the failure times T_1 and T_2 are potentially correlated. Wang and Ding (2000) examine their dependent relationship, and consider semi-parametric estimation of the association parameter in a bivariate copula model. Ding and Wang (2004) extend this work to develop a nonparametric inference procedure for testing independence between the two failure time variables, when only bivariate current status data is available. In a more recent paper, Wang et al. (2008) consider efficient estimation of regression and association parameters jointly for bivariate current status data, assuming a marginal proportional hazards model.

1.5 Counting Processes

Both the simple current status data structure, introduced in Section 1.1, and the more complex extensions of this framework can be viewed as a counting process. For a deeper understanding of counting processes, specifically in the survival analysis setting, see Fleming and Harrington (2005). For an overview of counting processes applied to current status data see Sun and Kalbfleisch (1993). Simple current status data, where there is a single survival time random variable of interest T, can be described directly by a simple counting process consisting of at most one jump occurring at the failure time T. This simple counting process can be defined by

$$N(C) = \begin{cases} 0 & \text{if} \quad C < T \\ 1 & \text{if} \quad T \leq C \end{cases} . \tag{1.2}$$

This simple counting process can be extended to allow for multiple consecutive events, or jumps, namely, multistate current status data, which is described in more detail in the next section. The special case of multistate current status data consisting of exactly two jumps, creating three states, is considered in Chapter 4, with an application to simultaneous accurate and diluted HIV test data. With two consecutive events T_1 and T_2, Chapter 4 considers nonparametric estimation of time to the first event, and specifically whether current status information on a subsequent event can be used to improve this estimate.

1.5.1 Multistate Current Status Data

Multistate models, as described by Anderson and Keiding (2002) and Meira-Machado et al. (2009), are a type of multivariate survival data, used to explain how individuals move through a succession of stages, corresponding to distinct states. The simplest example of a multistate model for survival data is the mortality model, made up of only two states, where individuals move from an initial stage (alive) to a terminal stage (dead). This corresponds to the simple counting process described in Equation (1.2) above, consisting of a single event of interest. In multistate current status data (Datta and Sundaram, 2006, Lan and Datta, 2008) one only observes whether or not each of the individual survival times defining the different stages exceed the common observed monitoring time. An important difference between multistate current status data and bivariate current status data is that in the former there is a specific ordering to the events, namely the first event must occur at or before the second event, and so on. Therefore, if the first event has not occurred by the monitoring time C, it is not possible that any subsequent event has occurred before this time. Conversely, if the final event of interest has occurred, then all preceding events must also have occurred.

When there are multiple ordered survival times, $T_1, T_2, ..., T_k$, a counting process is defined by $N(t) = \sum_{j=1}^{k} I(T_j \leq t)$, $T_1 < ... < T_k$, where T_j is the time variable at which an event occurs and the point where the count N jumps from $j-1$ to j. Different individuals can pass through a different number of states, or jumps, where the maximum number of potential jumps for any individual is denoted by k. Multistate current status data can be described as a sample of i.i.d. observations on the random variable $(N(C), C)$. The special case of multistate current status data with exactly three states, considered in Chapter 4, represents a counting process with the maximum attainable jumps fixed at $k = 2$. We then observe current status information on both the final and intermediate state.

1.6 Cure Models

A common assumption when analyzing survival data is that all individuals are initially subject to experiencing the event of interest. In many public health applications this assumption is appropriate where, at the beginning of the interval of interest, all individuals are at risk of developing a specific disease or experiencing a specific event. For example, at birth all individuals are at risk of death. However, this may not be true for all applications. For example, consider the demographic survey examining breastfeeding behavior, analyzed by Grummer-Strawn (1993). In this study interest is on estimating the time until weaning in developing countries. As is expected in such an example, some children are never breastfed and are therefore not subject to weaning at a later age. Grummer-Strawn (1993) therefore conduct the analysis in terms of the proportion of children ever breastfed. Similarly, it is commonly assumed that all individuals will eventually experience the event of interest. In mortality studies, this is of course an accurate assumption but examples also exist where this assumption no longer holds. In medical studies there may exist a subgroup of individuals who are not susceptible to the specific disease of interest. Therefore, a proportion of individuals will never experience the event of interest and consequently the survival curve will eventually level out without reaching a value of 0. Special survival analysis models have been developed for this type of data to estimate the proportion of subjects who do not experience the event. Such models have been referred to as cure models, under which there will always be at least some survivors, resulting in $S(\infty) > 0$. For more details on cure models see Maller and Zhou (2001).

As stated by Ma (2009), when considering current status data, previous methodological studies including nonparametric, semi-parametric and parametric models, assume that if the follow up time is long enough, all subjects will eventually experience the event of interest. These studies include the nonparametric model in Groeneboom and Wellner (1992), the linear Cox model in Huang and Wellner (1997), the additive risk model in Lin et al. (1998), and the partly linear accelerated failure time model in Ma and Kosorok (2005). These studies also assume that at the beginning of the time period of interest, all individuals are at risk of experiencing the event of interest. Ma (2009) consider current status data with a cured subgroup where individuals within that subgroup are not susceptible to the event. Cure models have been considered by other authors for related data structures such as right censored survival data (Fang et al., 2005, Chen et al., 2004, Lu and Ying, 2004) and interval censored data (Liu and Shen, 2009, Lam and Xue, 2005, Thompson and Chhikara, 2003). To address current status data with a cured subgroup Ma (2009) assume the cure probability satisfies a generalized linear model with a known link function. For susceptible individuals, the event time is modeled using linear or partly linear Cox models. A (penalized) maximum likelihood approach and weighted bootstrap are used for estimation and inference, respectively. In a recent paper, Ma (2011) also consider current status data with a cured subgroup, where the event time for susceptible individuals is modeled using a partly linear additive risk model. As a flexible alternative to the Cox model, the additive risk is adopted when the covariates contribute to the hazard function in an additive manner.

1.7 Current Status Data with Competing Risks

For simple current status data (Section 1.1) the failure time random variable T is assumed to have a single definition, for example, time of infection, time of death, etc. Although this is the most commonly used form of current status data, there are many examples in which there may exist multiple causes of failure. The failure time T can therefore have more than a single definition. An individual can experience only one of potentially many distinct causes of failure. When there are multiple sources of failure, an individual is subject to multiple risks, only one of which can occur. Data of this form are often referred to as competing risks data. As an example, consider estimating the time until menopause and note that menopause can either be natural or operative, only one of which can occur (Jewell et al., 2003). This is the simplest scenario with only two competing risks. The competing risks framework has been considered for both right censored data (Kalbfleish and Prentice, 2002) and general interval censored data (Hudgens et al., 2001).

Competing risks can also be applied to current status data. Extending the notation of Section 1.1 to incorporate the competing risks framework, let J be the random variable indicating the cause of the failure at time T. Jewell et al. (2003) consider estimation of the sub-survival functions, for each cause of failure, based on current status observations in the presence of competing risks. Assuming there are m distinct failure types, the sub-distribution functions of interest can be defined by

$$F_j(t) = P(T \leq t, J = j) \quad j = 1, 2, ..., m.$$

The overall survival function can then be given by

$$S(t) = 1 - F_1(t) - F_2(t) - .... - F_m(t).$$

For simplicity, Jewell et al. (2003) consider the case of only two competing risks but the ideas are easily extended to allow for the possibility of many competing causes of failure. For current status data with only two competing risks, the observed data can be described as (Y, C) with Y now defined as $Y = (\Delta, \Phi)$, where $\Delta = 1$ if $T \leq C$ with $j = 1$ and $\Phi = 1$ if $T \leq C$ with $J = 2$. Assuming the observed monitoring times are independent of T and are uninformative, the conditional likelihood for such a sample can be given by

$$CL = \prod_{i=1}^{n} F_1(c_i)^{\delta_i} F_2(c_i)^{\phi_i} S(c_i)^{1-\delta_i-\phi_i}.$$

Jewell et al. (2003) consider maximization of this likelihood for simple parametric models as well as nonparametric estimation through both an ad hoc (naïve) estimator and the nonparametric maximum likelihood estimator (NPMLE). We use the ideas of this ad hoc approach to develop a so called sophisticated naïve estimator in Section 4.2.1 in relation to current status observation of a three state counting process. This naïve estimator of Jewell et al. (2003) is not the NPMLE but they show that smooth functionals of either sub-distribution function are efficiently estimated using the appropriate functionals of the naïve estimators of F_1 or F_2. A faster algorithm for estimation of these NPMLEs which generalizes the pool-adjacent violators algorithm is provided by Jewell and Kalbfleisch (2004).

Groeneboom et al. (2008) and Maathuis (2011) develop on the results of this paper by deriving the large sample behaviors of both the naïve estimator and the maximum likelihood estimator. Groeneboom et al. (2008) assume a model that imposes certain smoothness conditions on F and the observation time distribution. It is shown that both estimates have the same local rate of convergence of $n^{1/3}$. Groeneboom et al. (2008) show that the limiting distribution of the maximum likelihood estimate is non-standard and involves a self induced system of slopes of convex minorants of Brownian motion processes plus parabolic drifts. The limiting distribution for the naïve estimator is simpler and does not involve a self-induced system. Maathuis (2011) examine the large sample behavior of both estimators under a discrete model and a model with grouped observation times. In the discrete model it is shown that both estimates have a local rate of convergence of $n^{1/2}$, with both limiting distributions being identical and normal.

1.8 Informative Censoring

A common assumption used with current status data is that the monitoring time is independent of the underlying survival time. In many data applications this is a reasonable assumption. However, certain examples exist where this assumption is perhaps no longer the most appropriate, and informative censoring, where there monitoring time may depend on the survival time of interest, could be preferred. Zhang et al. (2005) consider regression analysis of current status data when the observation time may be related to the underlying survival time. Inference procedures are presented for estimation of regression parameters under the assumption of an additive hazards regression model.

The existence of monitoring times that depend on the survival time of interest are not limited to the simple current status data scenario. Such a dependence is equally likely for the more complex extensions of current status data. Dunson and Dinse (2002) address the issue of informative censoring with multivariate current status data, a generalization of the bivariate current status data considered in Section 1.4.5. Dunson and Dinse (2002) propose a Bayesian model for the analysis of multivariate current status data subject to informative censoring. It is however assumed that the different events occur independently.

1.9 Current Status Data with Misclassification

Current status data are generally considered more reliable than retrospective reports of age or duration because respondents can more accurately report their current behavior rather than recall the timing of an event that occurred in the past (Lesthaeghe and Page, 1980). For this reason, in applications in demography, such as in the analysis of the duration of breastfeeding, current status observations are the primary form of data collected (Grummer-Strawn, 1993). However, as with any form of survival data, current status data may also be subject to misclassification of various forms. The current status outcome or the covariates in the regression setting may potentially be observed with error. In general, the monitoring time is assumed to be accurately observed.

1.9.1 Misclassified Outcomes

In many current status data applications, including those described in Section 1.1.1, ascertainment of an individual's current status is based on a screening test which may not have perfect sensitivity or specificity. For example, tests for the infection status of a viral disease like HIV or HPV are designed to detect antibodies and may be subject to error particularly when a test is performed soon after infection. Detection of the existence of uterine fibroids through ultrasounds (Young et al., 2008) is known to be subject to error. When an individual's current status is measured through a survey instrument, such as studies examining the age at weaning (Grummer-Strawn, 1993), or the age at onset of menopause (Jewell et al., 2003), there is potential for misclassification particularly close to the (unobserved) event time, menopause in this specific example.

Chapter 2 considers misclassification of current status data where the outcome variable Y is subject to misclassification. The nonparametric maximum likelihood estimator of the distribution function underlying current status data, when there is no misclassification, is extended to allow for time-independent misclassification of both apparent survivors and failures, using known misclassification rates. Calculation of the proposed estimator to allow for such misclassification uses a simple modification of the pool-adjacent violators algorithm. Details of this modification are given in Section 2.2. Asymptotic properties therefore follow straightforwardly. The implication of misclassification rates that vary over time is also considered, in particular when misclassification only occurs within a known time window surrounding the underlying failure event. Regression models are also considered where the current status outcomes are observed with error, using the ideas for binary generalized linear models with the outcome variable subject to misclassification (Neuhaus, 1999). Chapter 2 presents results and plots for simulated data along with an application to the HPV infection status of women in San Francisco, the details of this dataset are described within. Chapter 2 assumes known misclassification rates feasible for this specific example. Similar figures and tables are presented in Appendix A for two alternative data examples, using corresponding feasible classification rates.

Sal y Rosas and Hughes (2011) develop on the work of Chapter 2. In Chapter 2 we assume all individuals were tested using the same testing mechanism and are therefore subject to identical levels of sensitivity and specificity. Sal y Rosas and Hughes (2011) consider the case where individuals can be tested with different laboratory tests, where a proportion of individuals receive a more accurate test than the others. Hypothesis testing with two current status data samples (intervention and control groups) are considered, where both samples are subject to outcome misclassification. The regression model for current status data with the outcome variable subject to misclassification considered in Section 2.2.5 is also extended to consider a semi-parametric proportional hazards model.

1.9.2 Mismeasured Covariates

As discussed in Section 1.3. considerable attention has been given to the examination of the relationship between the failure time random variable and a set of covariates, assuming various regression models. However, most models considered assume that both the outcome variable and the covariates are observed without error. In practice, not only the outcome variable is subject to imperfect sensitivity or specificity (as described in Chapter 2), the covariates may also be observed with error. Several authors have considered mismeasured covariates in the context of right-censored survival data (Hu et al., 1998, Tsiatis and Davidson, 2001), and also for interval censored data (Song and Huang, 2005). However, these methods do not directly apply to the current status data framework considered here. Wen et al. (2011) describe a semi-parametric maximum likelihood method for analyzing current status data with mismeasured covariates under the proportional hazards model. They show that the estimator of the regression coefficient is asymptotically normal and efficient and that the profile likelihood ratio test is asymptotically Chi-squared. They also provide an algorithm for computing the estimators which can be easily implemented.

As with other forms of data, current status data can potentially be observed with missing covariates. Previous authors have considered statistical inference on right-censored data with missing covariates, under the proportional hazards assumption (Qi et al., 2005, Zhou and Pepe, 1995, Lin and Ying, 1993). In a recent paper, Wen and Lin (2011) extend this work and consider current status data with missing covariates, also assuming a proportional hazards model. The missing covariates are assumed to be missing at random, with the monitoring time assumed independent of the missing covariates. A semi-parametric maximum likelihood method is proposed to analyze the current status data in this scenario. An algorithm for computing the proposed semi-parametric maximum likelihood estimator, along with the variance estimate and profile likelihood ratio statistic are given.

1.9.3 Group Testing with Misclassification

The use of screening tests to detect the presence of an infectious disease is often limited by cost or other factors. To control the spread of such infectious diseases it is often necessary to test a large number of individuals. Bilder et al. (2010) consider the use of group testing (also known as batch testing or pooled testing) to reduce the number of required tests. Group testing combines samples, such as blood or urine, from a number of individuals and then tests the combined (group) sample rather than testing each individual sample for the presence of the disease of interest. The original group testing procedure of Dorfman (1943) involves testing the group sample, from which if the group tests negative, then all of the individual samples within that group are declared negative. If a group tests positive, then each sample within the group is subsequently tested individually to identify the positive samples. Many variations of this strategy were later developed using different pooling algorithms on the positive sample rather than retesting each individual sample, see for example Hughes-Oliver and Swallow (1994), Brookmeyer (1999), Bilder et al. (2010), Xie et al. (2001), amongst others.

Again, the testing mechanism used for group testing may not have perfect sensitivity or specificity. In Chapter 3 we consider group testing from a current status data perspective, specifically when the current status outcome is subject to misclassification. We extend the results of Chapter 2 and develop on the work of Liu et al. (2011) who consider estimation of disease prevalence, the probability of a positive test result at a single monitoring time. In the group testing framework, when a positive group is identified, retesting individuals to find those with a positive sample is not necessary for estimating this prevalence. We also consider determination of the optimal group size for estimation of this prevalence, under various scenarios.

1.10 Mortality Differentials

An underlying assumption of many current status data applications is that individuals for whom the event of interest has occurred are equally likely to be selected at the monitoring time as those for whom the event has not yet occurred. This is an important assumption which implies there are no mortality differentials between a positive and negative current status outcome at time C. There will be no mortality selection bias if the probability of mortality is the same for all individuals, regardless of their event status. This assumption is reasonable for many applications, including examining the time until first HPV infection, natural menopause or weaning, some of the applications addressed in this dissertation. However, in certain examples, mortality differentials may exist when an individual's mortality risk increases due to the presence of the disease of interest. For example, consider a study of HIV infection status. In this case, there may exist mortality selection bias when monitoring individuals at different ages. Individuals with and without HIV may be equally likely to be selected in a young age group. However, individuals who have had HIV for a long time, and are thus in an older age group, may die before the survey as a result of HIV infection, preventing them from providing information about their HIV status.

The remaining chapters of this dissertation assume no differential selection but more sophisticated models could be developed to allow for such a bias. In particular, when considering current status observation of a three-state counting process in Chapter 4 it is assumed that the probability of being selected at a particular monitoring time does not vary from state to state.

1.11 Summary

The remaining chapters of this dissertation add to the existing literature on current status data and the appropriate methods of analysis for use with this type of data. Existing well-known methods, such as the PAV algorithm of Section 1.2 are extended to incorporate concepts not previously addressed, such as misclassification of the response variable (Chapter 2). We then further extend this work to consider the case where instead of testing each sample individually, the samples are combined and examined as a group, as described in Section 1.9.3. We compare estimation of the distribution of time to a specific event for both the individual and group setting when the outcomes are subject to misclassification (Chapter 3), and determine the scenarios under which group testing is preferred to testing each sample individually.

As was described in Section 1.5, current status data, including the setup considered in Chapter 2 and Chapter 3, can be viewed as a counting process. In Chapter 4 we approach the analysis of current status data from a counting process perspective. Specifically, we consider current status observation of a three-state counting process, defined by two events of interest. We examine estimation of the time until the first event and whether current status information on the second event can be used to improve this estimate. Some applicable extensions and open problems in current status data are described at the end of each chapter, with some concluding remarks given in Chapter 5.

Chapter 2

Misclassification of Current Status Data

2.1 Introduction

Recall from Chapter 1 that current status data is a type of survival data that provides information on the survival status of individuals at various times rather than standard observation of the failure times. As described in Section 1.9, current status data may be subject to various forms of misclassification. This chapter specifically addresses the case where the current status outcome is subject to misclassification. Although the general concept of current status data and basic notation has been described in the Chapter 1, for completeness and ease of understanding, Section 2.2 introduces the necessary variables and notation used throughout this chapter. As was seen in Chapter 1, considerable attention has been given to estimation of a survival function based on both simple current status data, and its more complex extensions. Attention has also been given to estimation of regression coefficients from a variety of standard models, where the techniques of generalized linear models can be utilized. Current status data has been examined for many years, with some of the earliest work applied to areas in demography (Diamond et al., 1986) and epidemiology (Becker, 1989), followed by carcinogenicity studies (Gart et al., 1986), partner studies of Human Immunodeficiency Virus (HIV) transmission (Shiboski and Jewell, 1992), age-incidence estimation, and assessment of environmental exposures (Keiding, 1991). Nonparametric estimation in the single-sample setting is based on the well-known pool-adjacent violators (PAV) algorithm of Ayer et al. (1955), see Section 1.2 for more details on this algorithm. A review of various important aspects of current status data are given in Chapter 1.

In many commonly used current status applications, ascertainment of an individual's current status is based on a testing mechanism that may not have perfect sensitivity or specificity. Misclassification of a current status outcome may occur through the use of a survey instrument, or through laboratory error when testing infection status, or the presence of a specific antibody. Due to the nature of the infection, a test used to detect antibodies of a viral disease like HIV or HPV may be subject to error, particularly when a test is performed soon after infection. When obtaining current status data from a survey, such as in studies to examine the age at onset of menopause (Jewell et al., 2003), there is potential for misreporting due to the silent nature of the event, in this case, menopause. Again, imperfect classification may occur particularly close to the time of the unobserved event.

In this chapter we extend the PAV algorithm commonly used for nonparametric maximum likelihood estimation of the distribution function underlying current status data when there is no misclassification. The next section describes how this algorithm, using known misclassification rates, can be modified to allow for time-independent misclassification of both the apparent survivors and failures. Asymptotic properties of the proposed estimator follow straightforwardly from those of the standard PAV algorithm. In Section 2.2.3 we consider misclassification rates that vary over time. In particular, we consider the case where misclassification is most likely within a window surrounding the time of the true failure event, and in certain cases misclassification only occurs within this window, with perfect sensitivity and specificity observed elsewhere. Regression models for current status data subject to misclassification are also considered where the link between current status data and generalized linear models is described both in the presence and absence of this misclassification.

2.2 Nonparametric Estimation of a Single Distribution Function

Continuing the notation from Section 1.1.2, assume the standard data structure for simple current status data with the following notation. Let T be the survival time random variable of interest with distribution function F, and let the monitoring time be denoted by the random variable C. As usual, we assume that C is independent of T. As was seen in Section 1.8, in some examples, C is non-random. In either case, we focus directly on the conditional likelihood, given C. For convenience we describe the random monitoring time scenario, where current status observation refers to a sampling scheme with n i.i.d. observations collected on the random variable (Y, C) where $Y = I(T \leq C)$. Most of the ideas presented are based on the conditional likelihood, given C, and would therefore also be equally applicable for the non-random monitoring time scenario.

Motivated by the examples discussed in the Section 1.1.1 and the potential for misclassification, as highlighted in Section 1.9, we now consider the possibility that the random variable Y is observed with error. We focus primarily on the following constant misclassification model where the levels of sensitivity and specificity are assumed known, remain constant over time and are identical for all individuals, although we discuss alternative error models in Section 2.2.3. To address this form of misclassification, assume that instead of observing the current status outcome Y we now observe the random variable Δ where

$$P(\Delta = 1|Y = 1) = \alpha \qquad\qquad P(\Delta = 0|Y = 0) = \beta.$$

The classification rates α and β therefore define the sensitivity and specificity, respectively. The observed data can be described as n i.i.d. copies of (Δ, C). In Section 2.2.5, we also incorporate covariates (observed without error) for each individual.

We assume the true classification probabilities α, $\beta > 0.5$, and are the same for each individual. Sal y Rosas and Hughes (2011) later extend this work to consider the case where the testing mechanism, and corresponding classification rates, may differ between individuals. We also assume that these classification rates do not depend on the monitoring time. Continuing with our notation, let C_i be the i^{th} order statistic of $C_1, C_2, ..., C_n$. Given that the monitoring time C is independent of the survival time T, and Δ is independent of (C, T), under this misclassification model, the (conditional) likelihood function is given by

$$\prod_{i=1}^{n} [P(\Delta_i = 1|c_i)]^{\delta_i} [P(\Delta_i = 0|c_i)]^{1-\delta_i}, \tag{2.1}$$

where c_i is the observed value of C_i and δ_i is the observed value of Δ_i. The probabilities $P(\Delta_i = 1|c_i)$ and $P(\Delta_i = 0|c_i)$ can be defined as follows

$$\begin{aligned} P(\Delta_i = 1|c_i) &= P(\Delta_i = 1|y_i = 1, c_i)P(y_i = 1|c_i) + P(\Delta_i = 1|y_i = 0, c_i)P(y_i = 0|c_i) \\ &= (\alpha - 1 + \beta)F(c_i) + 1 - \beta, \end{aligned}$$

and

$$\begin{aligned} P(\Delta_i = 0|c_i) &= P(\Delta_i = 0|y_i = 0, c_i)P(y_i = 0|c_i) + P(\Delta_i = 0|y_i = 1, c_i)P(y_i = 1|c_i) \\ &= \beta - (\alpha - 1 + \beta)F(c_i). \end{aligned}$$

For ease of notation let $\gamma = \alpha + \beta - 1 > 0$. Then the (conditional) likelihood function allowing for constant misclassification in the response variable can be written as

$$\prod_{i=1}^{n} [\gamma F(c_i) + (1 - \beta)]^{\delta_i} [\beta - \gamma F(c_i)]^{1-\delta_i},$$

with corresponding log-likelihood given by

$$\sum_{i=1}^{n} \delta_i \log(\gamma F(c_i) + (1 - \beta)) + \sum_{i=1}^{n} (1 - \delta_i) \log(\beta - \gamma F(c_i)).$$

Writing $G(c_i) \equiv \gamma F(c_i) + (1-\beta)$, then the nonparametric maximum likelihood estimate of the distribution function when the current status outcomes are subject to misclassification can be found by obtaining a vector $\tilde{z} = (z_1 = \hat{G}(c_1), \ldots, z_n = \hat{G}(c_n)) \in R^n$ maximizing

$$\phi(G(c_i)) = \sum_{i=1}^{n} \delta_i log(G(c_i)) + \sum_{i=1}^{n} (1-\delta_i) log(1-G(c_i)) \tag{2.2}$$

under the constraint

$$1-\beta \leq G(c_1) \leq G(c_2) \leq \cdots \leq G(c_n) \leq \alpha. \tag{2.3}$$

Note that G is itself a distribution function, the values of which must be non-decreasing at successive monitoring times and fall between zero and one. We now formally show how modifications can be made to the PAV algorithm to obtain the nonparametric maximum likelihood estimate (NPMLE) of the distribution function F, when the current status outcomes are potentially observed with error. Throughout the remainder of this chapter, we refer to the NPMLE assuming no misclassification as the unconstrained NPMLE and the NPMLE allowing for misclassification as the constrained, or adjusted, NPMLE. This adjusted NPMLE assumes a constant rate of misclassification that does not depend on time.

Claim:
The identity $z_m = \min(\max(\hat{z}_m, 1-\beta), \alpha)$, with $m = 1, \ldots, n$, defines the unique vector $\tilde{z}$, where $\tilde{z} = (z_1, z_2, \ldots, z_n) \in \mathbb{R}^n$ maximizes Equation (2.2) under the constraint described in Equation (2.3), with $G(c_i)$ replaced by z_i, where

$$\hat{z}_m = \max_{i \leq m} \min_{k \geq m} \frac{\sum_{i \leq j \leq k} \delta_i}{k-i+1}$$

is the unconstrained NPMLE of the distribution function G based on the likelihood in Equation (2.2) but with no additional constraints.

Note that the vector $\{\hat{z}_m : m = 1, \ldots, n\}$, and consequently the estimate of G, can be computed using the standard pool-adjacent violators algorithm, originally described by Ayer et al. (1955) and characterised by Barlow et al. (1972) and Groeneboom and Wellner (1992) in terms of convex minorants. The vector $\{z_m\}$ modifies any value of $\{\hat{z}_m\}$ less than $1-\beta$ to equal $1-\beta$, and similarly modifies any value of $\{\hat{z}_m\}$ greater than α to equal α. The estimate of the NPMLE of F at a monitoring time c_i then follows using the relationship

$$\hat{F}(c_i) = [\hat{G}(c_i) - 1 + \beta]/\gamma. \tag{2.4}$$

Proof of Claim:
First note that, if $\delta_i = 0$ for $i = 1, 2, \ldots, k$, then maximizing Equation (2.2) requires the second term to be as large as possible, in which case, we set $z_1, z_2, \ldots, z_k = 1-\beta$ without affecting the maximization problem over the remaining $z_{k+1}, \ldots, z_n$. Similarly, if $\delta_i = 1$ for

$j \leq i \leq n$, then to maximize Equation (2.2) we make the first term as large as possible, setting $z_j, z_{j+1}, \ldots, z_n = \alpha$.

Suppose there exists at least one $\delta_i = 1$ followed by a $\delta_j = 0$, for some $j > i$ (otherwise we are done).

Let k_0 be the smallest index i such that $\delta_i = 1$, and let k_1 be the smallest index $k \geq k_0$ such that

$$\max_{i \leq m} \min_{k \geq m} \frac{\sum_{i \leq j \leq k} \delta_i}{k - i + 1} \geq 1 - \beta.$$

Analogously, let m_0 be the largest index $k \geq k_1$ such that

$$\max_{i \leq m} \min_{k \geq m} \frac{\sum_{i \leq j \leq k} \delta_i}{k - i + 1} \leq \alpha,$$

with m_1 being the largest index i such that $\delta_i = 0$.

Thus, k_0 and m_1 represent the index of the first $\delta_i = 1$ and the last $\delta_j = 0$ respectively, where $j > i$. Also, $k_1 - 1$ is the smallest index for which the unconstrained NPMLE does not fall below $1 - \beta$, and m_0 is the largest index for which the unconstrained NPMLE does not go above α. Figure 2.1 shows the positioning of such indices as they would appear in terms of a hypothetical unconstrained NPMLE of a distribution function. The dashed lines are positioned at $1 - \beta$ and α, between which the constrained NPMLE must lie.

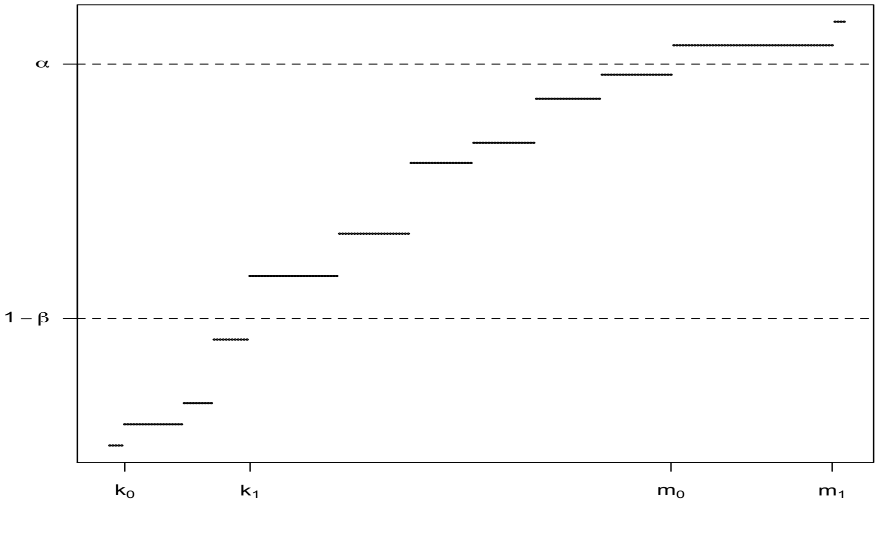

Figure 2.1: Schematic of a hypothetical unconstrained NPMLE of a distribution function, showing the values between which the corresponding adjusted NPMLE must lie.

Using the above definitions of the required indices, we can redefine the claim by the following three statements

(A) For all indices $m < k_1$, $z_m = 1 - \beta$.

(B) For all indices $m > m_0$, $z_m = \alpha$.

(C) For all indices $k_1 \leq m \leq m_0$, $z_m = \max_{i \leq m} \min_{k \geq m} \frac{\sum_{i \leq j \leq k} \delta_i}{k-i+1}$, the unconstrained NPMLE.

We prove the claim by establishing each statement separately. First, we show that for all indices $m < k_1$, $z_m = 1-\beta$ maximizes the relevant terms in the likelihood of Equation (2.2), subject to the constraint given in Equation (2.3) without affecting the optimization function, or constraint, based on z_i for other indices. A detailed proof is given for statement (A), from which proofs of statements (B) and (C) follow straightforwardly.

Proof of Statement (A):
Consider indices i for $k_0 \leq i < k_1$. Suppose the values of z_i over this range of indices take values that are increasing and, are necessarily greater than $1 - \beta$.

Take the largest of these indices (just to the "left" of k_1) where the proposed maximizer values of z_i assume the value $1 - \beta + \epsilon$, with $\epsilon > 0$. It does not matter here whether z_i assumes this value at one index, or over a set, $\tilde{S}$ of consecutive indices.

Assume that amongst the set of indices, $\tilde{S}$, there are p indices where $\delta_i = 1$ and q indices where $\delta_i = 0$. The contribution to the likelihood in Equation (2.2) over this set of indices is therefore, say,

$$p \log(1 - \beta + \epsilon) + q \log(\beta - \epsilon) \equiv h(\epsilon).$$

The derivative of this function is

$$h^{'}(\epsilon) = [p/(1 - \beta + \epsilon)] - [q/(\beta - \epsilon)].$$

Now, by the definition of k_1, relative to the definition of the unconstrained NPMLE, it follows that $p/(p+q) < 1 - \beta$. This in turn implies that $q/p > \beta/(1-\beta)$.

Since $\epsilon > 0$, it follows that $\beta/(1-\beta) > (\beta - \epsilon)/(1 - \beta + \epsilon)$, and $h^{'}(\epsilon) < 0$ so that h is decreasing in ϵ.

Thus, without changing the optimization problem in terms of the other indices and constraints, we can increase the likelihood by lowering the value of the proposed z_i to the next lower value (to the right) where $z_j = 1 - \beta + \lambda$ with $0 < \lambda < \epsilon$. However, we can now repeat the same argument in terms of λ, and thus we keep lowering the relevant z_i's until they all equal $1 - \beta$. This proves Statement (A).

An identical argument also establishes Statement (B).
The proof of Statement (C) then follows since $z_m = \max_{i \leq m} \min_{k \geq m} \frac{\sum_{i \leq j \leq k} \delta_i}{k-i+1}$ is already the unconstrained NMPLE and meets the constraints of Equation (2.3) by the definition of k_1 and m_0. The claim is thus proven.

2.2.1 Pointwise Confidence Intervals for the NPMLE

As was addressed in Section 1.2.1, the asymptotic properties of the standard NPMLE of F for current status data with no misclassification are non-standard. This topic has received considerable attention, where it has been shown that there is a slower rate of convergence ($n^{1/3}$ as opposed to the familiar $\sqrt{n}$ rate), and the limit distribution is not Gaussian (Groeneboom and Wellner, 1992). We conjecture straightforward extensions of these results for the NPMLE for misclassified current status data and therefore do not focus on the (asymptotic) variance of the NPMLE as a step towards confidence interval construction. In the case of perfect classification, various approaches have been developed for estimation of pointwise confidence intervals for F, using standard current status data (Banerjee and Wellner, 2005). Suggested techniques include the likelihood-ratio method (Banerjee and Wellner, 2001) and the m out of n bootstrap (Politis et al., 1999), both of which can presumably be adapted to allow for misclassification.

The standard bootstrap (Efron and Tibshirani, 1994) is known to be consistent in many situations, however Bickel et al. (1997) give many important examples where this bootstrap method may fail. In general, regardless of whether the data is sampled with replacement from the original current status data or generated from the NPMLE estimator, the standard bootstrap yields inconsistent estimates of the pointwise confidence intervals (Sen et al., 2010). Modifications to the standard bootstrap method, including a smoothed version of the bootstrap and the m out of n bootstrap, are appropriate in this setting (Politis et al., 1999). We use this m out of n bootstrap for confidence interval estimation in Section 2.2.2. For a more detailed description of the m out of n bootstrap see Bickel et al. (1997), Politis et al. (1999) and Santana (2009), amongst others.

The m out of n bootstrap is a variation of the the well-known bootstrap method where instead of resampling n observations from a sample of size n, we now resample fewer observations, $m < n$. Practically, this procedure involves choice of the 'block' size m. Asymptotically, m must be chosen so that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$. This approach can be used either with or without replacement. However, these requirements provide little guidance for a finite sample size. The choice of block size m is an important consideration. Banerjee and Wellner (2005) suggest an intricate procedure for choice of m, based itself on bootstrapping. Bickel and Sakov (2008) propose a data dependent rule for selecting m, and define when the standard bootstrap fails. The m out of n bootstrap method can be adapted to provide symmetric confidence intervals as these often perform better in finite samples. Banerjee and Wellner (2005) provide further implementation details. For current status data with misclassification, illustrative calculations of symmetric confidence intervals using the m out of n bootstrap are provided in Section 2.2.2. We do not however seek to obtain the optimal choice of m, instead we present the results for a variety of values of m, based on those examined in the literature.

2.2.2 Illustration and Data Example

This section addresses estimation of the NPMLE of F for both simulated data and an application to Human Papillomavirus (HPV) infection status. Similar analyses and corresponding plots were performed for two other current status data applications, the results of which are presented in Appendix A.

Human Papillomavirus Example

Current status data on human papillomavirus (HPV) infection among women motivate and illustrate this chapter. The dataset we analyze consists of 827 women examined in San Francisco, aged between 13.5 and 24.2 years. This data comes from a larger longitudinal study where we extract the current status information based on an individual's first examination. Therefore, the sample size of $n = 827$ can potentially differ to the sample size addressed by other authors with a similar version of this dataset. For more information on the testing mechanism used to determine HPV infection status and the way in which women were recruited for the study, see Moscicki et al. (1998). The data made available to us contained a binary indicator of whether a woman has HPV infection at the time of the survey (Y) and her age at screening (C). Covariates included indicators of current smoking status and past infection with any other sexually transmitted disease (STD). For more information about the dataset and the error rates determined for such tests, see Neuhaus (1999) where it was assumed that HPV testing approach enjoyed (correct) classification rates of $\alpha = 0.8$ and $\beta = 0.9$. We note that more advanced screening instruments for HPV are now available and that these classification rates may no longer be the most appropriate when examining HPV infection status under more advanced testing mechanisms. An overview of the human papillomavirus is given in Appendix B. This is not a comprehensive description of the virus but gives information on prevalence, biological features, detection and prevention, as well as a description of warts and cancer which can develop after being infected with HPV. Knowledge of these aspects of the viral infection are not necessary for understanding the data application of this chapter, instead Appendix B aims to highlight the importance of studying this infection, and the challenges faced in doing so.

As is described in Appendix B, the nature of HPV infection is such that in 90% of cases, the body's immune system clears HPV infection naturally within two years without any medical intervention (CDC, 2011b). Therefore, in this example we first need to consider the definition of the underlying failure time since HPV infection can go into remission in the sense that a negative test result can plausibly follow an earlier true positive test result. Here we define T to be age at *first* HPV infection as distinct from the cross-sectional prevalence interpretation used by Neuhaus (1999). In this case, we allow for additional misclassification of the apparently negative screening results as such individuals may previously have been infected, although they now have a true negative result. We assume that misclassification of this form applies to 10% of the negative screening results. This additional misclassification reduces the value of α from $\alpha = 0.8$ to $\alpha = 0.73$. This adjustment can be seen as follows

$$\begin{aligned}\alpha &= P(\Delta = 1|Y = 1)\\ &= P(\Delta = 1|Z = 1, Y = 1)P(Z = 1|Y = 1) + P(\Delta = 1|Z = 0, Y = 1)P(Z = 0|Y = 1),\end{aligned}$$

where $Z = 1$ if individual has antibodies. For the remainder of this section we therefore assume constant classification rates of $\alpha = 0.73$ and $\beta = 0.9$. We reconsider these classification rates in Section 2.2.3 when addressing misclassification that varies with time.

Unconstrained and Adjusted NPMLEs

Figure 2.2(a) illustrates the unconstrained NPMLE and the NPMLE adjusted for misclassification for a hypothetical dataset with sample size $n = 500$ generated from an Exponential distribution, F, with mean 2. The monitoring times were selected at random from a uniform distribution on a set of discrete time values ranging from 0 to 3 at equal increments of 0.1. The classification rates used in generating the data were $\alpha = \beta = 0.8$, and these values were assumed known in calculating the adjusted NPMLE. Note that, with $\alpha = \beta$, the two estimators cross at $\hat{F} = 0.5$, the estimated median time to occurrence. For earlier monitoring times, the estimate of F adjusted for misclassification is shifted downwards from the naïve, unconstrained estimator, as misclassifications are accounted for. Similarly, the adjusted NPMLE is shifted upwards at values of the monitoring time above the estimated median. Therefore, at the early monitoring times, values of the adjusted NPMLE are less than the corresponding values of the unconstrained NPMLE, with the reverse seen at the later times.

Figure 2.2(b) illustrates the HPV infection example described above. This figure displays both the unconstrained NPMLE of age at onset of first HPV infection, and the NPMLE adjusted for misclassification with the assumed values of $\alpha = 0.73$ and $\beta = 0.9$, allowing for the additional misclassification discussed above. Again it is assumed that these values of α and β are known (and remain constant) when calculating the adjusted NPMLE. With these unequal classification probabilities, the two curves cross at $\hat{F} = 0.270$, with the adjusted NPMLE shifted appropriately higher for higher ages. We do not see the shift downwards for lower ages since the first jump of the unconstrained NPMLE is to a value higher than 0.270. The additional data applications considered in Appendix A give examples where the two curves intersect at a value greater than the median time to occurrence, and also consider an example where there is perfect specificity, resulting in the curves intersecting at $\widehat{F} = 0$.

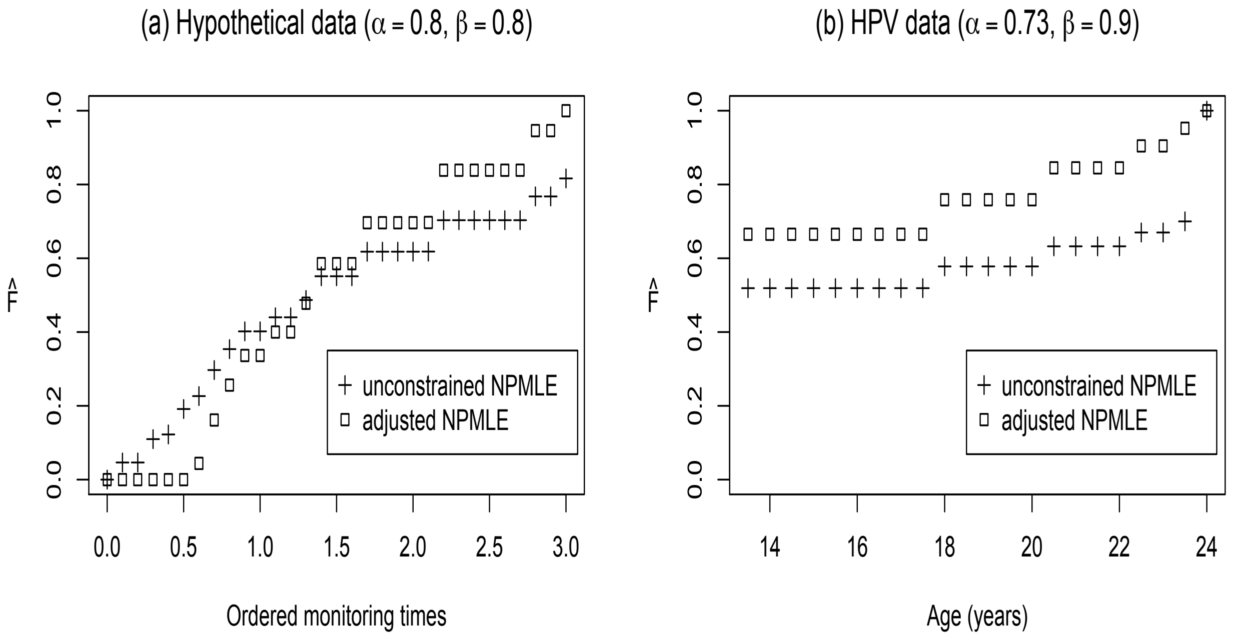

Figure 2.2: Estimated cumulative distribution functions for (a) hypothetical data and (b) HPV data. Both the unconstrained NPMLE obtained through the pool-adjacent violators algorithm and the proposed adjusted NPMLE are presented in each plot.

Symmetric Confidence Intervals

Using the m out of n bootstrap, described in Section 2.2.1, we estimate the 95% symmetric confidence intervals for the adjusted NPMLE of the distribution function of time to first HPV infection. Classification rates of $\alpha = 0.73$ and $\beta = 0.9$ are again assumed. Instead of searching for the optimal choice of block size m, we examine various choices of m ranging from 9 to 423, corresponding to values of m from $n^{1/3}$ to $n^{0.9}$. These block size values, m, were chosen based on the block sizes implemented in the simulations of Politis et al. (1999).

Table 2.1 provides the results of such calculations at three monitoring times, chosen to represent the spread of monitoring times. The results are quite stable across these choices except perhaps at 15.3 years. It it noteworthy that 15.3 years is close to the smallest monitoring times and is equal to the value of the first jump of the estimator. In Table 2.1 slightly more variability is suggested for the central values of m. However, overall the results provide a clear, consistent and useful assessment of variability. Similar estimates of

95% confidence intervals were calculated for the additional datasets, the results of which are presented in Appendix A.

t_0 $\hat{F}(t_0)$	15.3 years 0.652	19 years 0.739	22 years 0.829
$m = 9$ ($n^{1/3}$)	[0.503 0.801]	[0.569 0.909]	[0.723 0.935]
$m = 29$ ($n^{1/2}$)	[0.428 0.876]	[0.590 0.888]	[0.723 0.935]
$m = 88$ ($n^{2/3}$)	[0.333 0.971]	[0.622 0.856]	[0.744 0.914]
$m = 154$ ($n^{3/4}$)	[0.311 0.993]	[0.633 0.845]	[0.723 0.935]
$m = 216$ ($n^{0.8}$)	[0.407 0.897]	[0.633 0.845]	[0.712 0.946]
$m = 423$ ($n^{0.9}$)	[0.471 0.833]	[0.654 0.824]	[0.712 0.946]

Table 2.1: Confidence interval estimates for the adjusted ($\alpha = 0.73$, $\beta = 0.9$) NPMLE at three monitoring times obtained using the m out of n bootstrap for various values of m from 9 to 423.

2.2.3 Time-varying Misclassification

We now consider an extension of the simple, constant (i.e. time independent) misclassification model, introduced in Section 2.2, to allow the misclassification rates to vary over time. In particular, we consider the situation where the level of misclassification is defined by a known window surrounding the time of true event occurrence. This means that misclassification (both in terms of sensitivity and specificity) is higher when the observed monitoring time is close to the timing of the true event. A lower level of misclassification may exist outside this window, and in certain cases perfect classification may occur for monitoring times outside this window. This is natural for screening tests where accuracy may be improved, or even essentially perfect, far from the event time, either before or after the event. Misclassification is assumed more likely when the screening test is administered shortly before or after the event of interest. For example, when testing for a viral disease, a false negative is likely if an individual is examined immediately after infection as there are not yet enough detectable antibodies for a positive result. Therefore, in diagnosing HPV infection, the probability of a false negative possibly decreases with time since infection. Similarly, when considering current status information obtained through surveys, there is also the potential for misclassification close to the time of event occurrence, with perfect classification elsewhere. For example, with a current status assessment of menopause, misclassification is unlikely for a woman of age 30 or 65, but may be plausible at age 50. This data example is considered in Appendix A.

First, we examine the simple extension where misclassification occurs only within a time window surrounding the true failure event T given by $[T - A, T + A]$. Within this interval we assume the classification rates α, $\beta > 0.5$ are known, that perfect classification occurs at screening times outside this window, and that the value A is also known. Later in this section we see that a straightforward extension of this scenario allows all individuals to be subject to misclassification but at different levels defined by whether the monitoring time falls within A of T. For the basic time varying case, given the above assumptions, we define the log-likelihood as follows

$$\sum_{i=1}^{n} \delta_i log((1-\alpha)F(c_i - A) + (\alpha - (1-\beta))F(c_i) + (1-\beta)F(c_i + A))$$
$$+ \sum_{i=1}^{n} (1-\delta_i) log(1 - ((1-\alpha)F(c_i - A) + (\alpha - (1-\beta))F(c_i) + (1-\beta)F(c_i + A))). \quad (2.5)$$

Note, when $A = 0$ and $A = \infty$, Equation (2.5) reduces to the conditional log-likelihood of the unconstrained NPMLE and the conditional log-likelihood with constant misclassification rates (Equation (2.1)), respectively. This more complex conditional log-likelihood can still be written in the general form given in Equation (2.2) if we define a distribution function as

$$G^*(c_i) \equiv (1-\alpha)F(c_i - A) + (\alpha + \beta - 1)F(c_i) + (1-\beta)F(c_i + A).$$

However, finding the NPMLE of G^* is complicated here by the fact that the constraint on G^* (as $c \to 0$) depends on the unknown value of $F(A)$. In addition, even if a reasonable estimator of G^* is determined, it is not generally possible to solve directly for F in terms of G^*, as unlike in the time independent scenario (Equation (2.4)), a straightforward definition for F cannot be given in terms of known or estimated variables. This identifiability issue is most easily seen when there is but a single monitoring time, C. In this situation, only $G^*(C)$ is identifiable from the data and differing values of $F(C)$ (and $F(C-A)$ and $F(C+A)$) are compatible with any given value for $G^*(C)$. However, this does not address identifiability when the observed monitoring times cover a much broader range. In the latter situation, it is possible to make bias modifications to either the unconstrained or adjusted NPMLE to address an incorrect misclassification assumption. This allows the proposed and unconstrained estimators to accommodate a different window of misclassification than that assumed by either estimator. This bias-adjusted approach is formally introduced, discussed and evaluated via simulations in the next subsection.

2.2.4 Time-varying Misclassification: Simulations

To address the situation where misclassification is no longer independent of time, we carried out a set of simulations to further understand the implications of misclassification rates that vary over time. As there is no direct method available to calculate the NPMLE for this type of time-varying misclassification, we consider which of the two available estimators (the unconstrained NPMLE or the adjusted NPMLE) would be preferred if the data were truly misclassified in the manner described above. For these simulations, datasets of unobserved event times T, of sample size $n = 500$, were generated from an Exponential distribution, F, with mean 2. Current status observations were then created based on monitoring times, C, selected at random from a Uniform distribution on a set of discrete time values ranging from 0 to 3 at equal increments of 0.2. Finally, the current status data were (mis)classified with classification probabilities of $\alpha = \beta = 0.8$ if and only if $|C_i - T_i| \leq A$ in order to obtain the dataset used for estimation of F. Therefore, all individuals for whom the monitoring time falls within A of T are subject to misclassification, only some of which will actually be misclassified. Outside this window the current status responses were observed without error. A variety of values of A were examined including $A = 0$ (no misclassification) and $A = \infty$ (constant misclassification).

For each dataset, estimates of $\widehat{F}$ were obtained according to both the unconstrained NPMLE using the PAV algorithm directly, and the proposed estimator of Section 2.2 that assumes constant misclassification rates at all times and for all individuals (i.e. assumes $A = \infty$). For non-extreme values of A, where neither estimator directly represents the scenario we compare these two available estimators to determine which approach would be most accurate if it is suspected that the data is misclassified within a specific window and not misclassified otherwise. These simulations aim to highlight the impact of ignoring misclassification in estimating the time until a specific event of interest. Each simulation consisted of 1000 datasets.

C		0.4	0.8	1.4	1.8	2.8
		$F(C)$ =0.181	$F(C)$ = 0.330	$F(C)$ = 0.503	$F(C)$ = 0.593	$F(C)$ = 0.753
$A = 0$	0% (0)%					
	NPMLE$_0$	0.178(0.055)	0.331(0.063)	0.496(0.059)	0.591(0.056)	0.760(0.049)
	NPMLE$_\infty$	0.022(0.043)	0.218(0.104)	0.494(0.098)	0.652(0.094)	0.923(0.068)
$A = \infty$	100% (20)%					
	NPMLE$_0$	0.306(0.056)	0.397(0.056)	0.500(0.051)	0.557(0.047)	0.662(0.050)
	NPMLE$_\infty$	0.178(0.091)	0.329(0.094)	0.500(0.086)	0.593(0.078)	0.769(0.084)

Table 2.2: Simulation averages (standard deviations) of two estimators of the distribution function F (Exponential with mean 2) at 5 monitoring times, when the data generating distribution is either subject to always being misclassified ($A = \infty$), or never being misclassified ($A = 0$). The resulting percentage subject to misclassification (average percentage actually misclassified) are also given for each simulation. NPMLE$_0$ and NPMLE$_\infty$ represent the unconstrained NPMLE and the NPMLE adjusted for constant response misclassification, respectively.

Table 2.2 shows the results of both estimators of F at a selection of monitoring times, chosen systematically to depict the overall spread. These monitoring times are evaluated assuming windows of length $A = 0$ and $A = \infty$. The true value of F at each of the selected monitoring times is presented in the table, along with the percentage of individuals subject to misclassification (average percentage actually misclassified). In Table 2.2, and other tables to follow, NPMLE$_0$ represents the unconstrained NPMLE and NPMLE$_\infty$ represents the NPMLE adjusted for constant misclassification, at the appropriate constant levels. The results are as expected where the NPMLE assuming no misclassification performs best for a window of $A = 0$ (where no individuals are subject to misclassification) and the proposed NPMLE, adjusted for constant misclassification, performs best for a window of $A = \infty$ (where all individuals are subject to misclassification with an average of 20% truly misclassified).

C		0.4	0.8	1.4	1.8	2.8
		$F(C)$ $= 0.181$	$F(C)$ $= 0.330$	$F(C)$ $= 0.503$	$F(C)$ $= 0.593$	$F(C)$ $= 0.753$
$A = 1.5$	60%(12%)					
	NPMLE$_0$	0.225(0.057)	0.327(0.057)	0.454(0.056)	0.543(0.057)	0.732(0.052)
	NPMLE$_\infty$	0.063(0.069)	0.214(0.095)	0.424(0.094)	0.571(0.095)	0.884(0.079)
Bias	NPMLE$_0$	0.177(0.088)	0.315(0.097)	0.467(0.101)	0.575(0.102)	0.775(0.084)
adjusted	NPMLE$_\infty$	0.172(0.103)	0.337(0.116)	0.485(0.118)	0.588(0.119)	0.769(0.104)
$A = 2.5$	82%(16%)					
	NPMLE$_0$	0.258(0.055)	0.358(0.056)	0.470(0.056)	0.530(0.057)	0.673(0.051)
	NPMLE$_\infty$	0.104(0.084)	0.263(0.095)	0.451(0.087)	0.549(0.085)	0.787(0.090)
Bias	NPMLE$_0$	0.185(0.093)	0.320(0.099)	0.487(0.092)	0.551(0.090)	0.729(0.094)
adjusted	NPMLE$_\infty$	0.184(0.109)	0.331(0.117)	0.506(0.107)	0.559(0.105)	0.731(0.109)

Table 2.3: Simulation averages (standard deviations) of two estimators of the distribution function F (Exponential with mean 2) at 5 monitoring times when the data generating distribution is subject to constant misclassification ($\alpha = 0.8$, $\beta = 0.8$) only within a window of length $2A$ around the underlying failure time. Window lengths of $A = 1.5$ and $A = 2.5$ are evaluated. The resulting percentage subject to misclassification (average percentage misclassified) are also given for each simulation. NPMLE$_0$ and NPMLE $_\infty$ represent the unconstrained NPMLE and the NPMLE adjusted for constant response misclassification, respectively. The corresponding bias adjusted estimates (standard deviations) for each estimator under the different window lengths are also presented.

Table 2.3 provides similar results to those of Table 2.2, presented for the same monitoring times. However, in Table 2.3 the window length varies and the values of A do not represent either the unconstrained or adjusted likelihoods. In this table, for values of $A = 1.5$ and $A = 2.5$ approximately 60% and 82% of individuals are subject to misclassification, with an average of 12% and 16% actually being misclassified, respectively. The results of Table 2.3 are perhaps not as expected where the adjusted NPMLE only outperforms the unconstrained NPMLE when a very high proportion of individuals are subject to misclassification. Even when 82% of individuals are subject to misclassification, evidence in favor of the adjusted NPMLE is not overwhelming. There is also a visible bias with both estimators. The bias adjusted estimates given in Table 2.3 are calculated to try adjust the estimates to account for this bias. This issue is introduced and more formally addressed below, at which time we return to interpret the bias adjusted estimates in this table.

Bias Adjustment

In practice, an investigator necessarily does not know the underlying distribution F and so cannot immediately assess which approach to nonparametric estimation to use, that assuming no misclassification or assuming a constant rate of misclassification over time. In this situation, it is possible however to carry out a simulation using either estimator as the assumed 'true' F to examine performance. We also use this simulation to reduce the visible bias seen in Table 2.2 and Table 2.3. We apply this additional analysis to the previous simulations, and to the next set of simulations where we slightly increase the complexity of the misclassification model by allowing misclassification to exist outside the defined window, but at a reduced rate. Of course the varying levels of misclassification, defined by the window length, will not impact the cases where all ($A = \infty$)or none ($A = 0$) of the individuals are subject to the higher rate of misclassification.

If there is misclassification due to laboratory error in the (current status) screening instrument, all individuals will be subject to this error. However, even with constant laboratory misclassification, there may also be increased (and potentially asymmetric) misclassification rates close to the true failure event, for reasons outlined in Section 2.2.3. Table 2.4 presents results of simulations from the HPV data example where the true underlying distribution is assumed to be the unconstrained NPMLE, as obtained through the standard pool-adjacent violators algorithm. A constant laboratory error is assumed, giving classification rates of $\alpha = 0.8$ and $\beta = 0.9$ outside the window and $\alpha = 0.73$ and $\beta = 0.9$ within the window, indicating an additional deterioration in sensitivity close to the underlying failure time. In computing the adjusted NPMLE constant rates of classification are assumed with $\alpha = 0.73$ and $\beta = 0.9$, that is the adjusted NPMLE assumes all individuals are subject to misclassification at the highest error rate.

C		15 years	16.2 years	19.2 years	21.7 years	23.2 years
		$F(C)$ $= 0.484$	$F(C)$ $= 0.539$	$F(C)$ $= 0.581$	$F(C)$ $= 0.600$	$F(C)$ $= 0.661$
$A = 0$	0% (0%)					
	NPMLE$_0$	0.414(0.086)	0.464(0.054)	0.511(0.031)	0.540(0.039)	0.584(0.069)
	NPMLE$_\infty$	0.498(0.137)	0.578(0.085)	0.652(0.049)	0.698(0.061)	0.766(0.102)
$A = 4.5$	43%(7%)					
	NPMLE$_0$	0.388(0.084)	0.433(0.057)	0.498(0.034)	0.532(0.042)	0.584(0.076)
	NPMLE$_\infty$	0.457(0.132)	0.528(0.090)	0.632(0.054)	0.686(0.067)	0.764(0.111)
Bias	NPMLE$_0$	0.435(0.153)	0.475(0.113)	0.537(0.100)	0.562(0.115)	0.612(0.168)
adjusted	NPMLE$_\infty$	0.475(0.187)	0.472(0.149)	0.524(0.141)	0.544(0.147)	0.606(0.191)
$A = 8$	86%(15%)					
	NPMLE$_0$	0.383(0.081)	0.426(0.054)	0.474(0.030)	0.513(0.045)	0.573(0.078)
	NPMLE$_\infty$	0.449(0.129)	0.517(0.086)	0.594(0.047)	0.655(0.071)	0.747(0.114)
Bias	NPMLE$_0$	0.437(0.153)	0.473(0.113)	0.520(0.090)	0.553(0.114)	0.610(0.172)
adjusted	NPMLE$_\infty$	0.473(0.189)	0.472(0.150)	0.513(0.132)	0.542(0.154)	0.604(0.203)
$A = \infty$	100%(20%)					
	NPMLE$_0$	0.384(0.083)	0.428(0.052)	0.472(0.032)	0.496(0.038)	0.544(0.075)
	NPMLE$_\infty$	0.451(0.129)	0.521(0.083)	0.590(0.050)	0.629(0.060)	0.702(0.111)

Table 2.4: Simulation averages (standard deviations) of two estimators of the distribution function F (unconstrained NPMLE from the HPV data) at 5 monitoring times when the data generating distribution is subject to misclassification that varies with time. Classication rates of $\alpha = 0.8$ and $\beta = 0.9$ are assumed outside the window and rates of $\alpha = 0.73$ and $\beta = 0.9$ are assumed within a window of length $2A$ around the underlying failure time. Window lengths of $A = 0, 4.5, 8, \infty$ are evaluated. The resulting percentage subject to misclassification (average percentage misclassified) are also given for each simulation. NPMLE$_0$ and NPMLE $_\infty$ represent the unconstrained NPMLE and the NPMLE adjusted for constant ($\alpha = 0.73$, $\beta = 0.9$) misclassification, respectively. The corresponding bias adjusted estimates (standard deviations) for each estimator in the windows of length $A = 4.5$ and $A = 8$ are also presented.

In the simulations for the HPV infection status example, it must be noted that unlike the simulations in Table 2.2 and Table 2.3, when $A = 0$ there is still misclassification present, at a constant rate of $\alpha = 0.8$, $\beta = 0.9$. This explains the lack of accuracy in the unconstrained NPMLE for $A = 0$ which assumes no misclassification (and similarly for the constant misclassification adjusted NPMLE which uses the incorrect misclassification probabilities). When $A = \infty$ the results are as expected with the adjusted NPMLE more favorable, assuming constant misclassification at rates of $\alpha = 0.73$ and $\beta = 0.9$. Under the intermediate situations, with complex window misclassifications assuming non-zero and finite values for A, the simulations suggest that there is a slight preference for the adjusted NPMLE in terms of bias although there is a small price to be paid for additional variability. The mean squared error here gives a slight preferential tendency toward the unconstrained NPMLE, at least with these two possibilities for the window parameter A.

Regardless of the preferred approach for estimation of the NPMLE of F, the simulations suggest a way to remove the bias for either estimator. The approach can be used for all values of A, though only the results for finite and non-zero values of A are presented in the tables. The bias-adjusted algorithm we propose is as follows

(i) Compute a suitable simulation 'guess' for F to be used in the simulations.
(ii) Simulate data assuming this 'guess' is the truth, with the assumed value for A and the relevant misclassification probabilities within and outside the window defined by A.
(iii) Estimate the bias at all monitoring times of interest by comparing the simulation average with either of the original estimators (NPMLE$_0$ or NPMLE$_\infty$).
(iv) Remove this estimated bias from the original estimator.

Either the unconstrained NPMLE or the NPMLE adjusted for constant misclassification could be used for the initial 'guess', although the average of these two straightforward estimators may be preferred since the simulations seem to suggest that the bias for the two estimators is sometimes in opposite directions, particularly in the tails where the biases tend to be most severe. Note that this algorithm can also be used for more complex misclassification models that might be anticipated.

We now formalize the above steps of the bias-adjusted approach.
Define the bias in the unconstrained NPMLE at time t_0 as

$$bias_0(t_0) = E(\hat{F}_0(t_0, F)) - F(t_0),$$

where F is the assumed true data generating distribution, and $\hat{F}_0$ is the unconstrained NPMLE. We estimate the bias by substituting $\hat{F}_g$ for F in each of the terms in $bias_0(t_0)$ and estimate the expectation through simulations, thus yielding

$$\hat{bias}_0(t_0) = \hat{E}(\hat{F}_0(t_0, \hat{F}_g)) - \hat{F}_g(t_0).$$

This estimate, $\hat{F}_g$, could be the unconstrained estimate, $\hat{F}(t_0, F) = \hat{F}_0(t_0, F)$, the estimate under constant misclassification, $\hat{F}(t_0, F) = \hat{F}_\infty(t_0, F)$, or the average of both estimates,

$\hat{F}(t_0, F) = (\hat{F}_0(t_0, F) + \hat{F}_\infty(t_0, F))/2$. Finally we produce the bias-adjusted estimate for the unconstrained NPMLE as

$$\hat{F}_0^{ba}(t_0) = \hat{F}_0(t_0, F) - \hat{bias}_0(t_0).$$

Similarly, define the estimated bias for the adjusted NPMLE as

$$\hat{bias}_\infty(t_0) = \hat{E}(\hat{F}_\infty(t_0, \hat{F}_g)) - \hat{F}_g(t_0),$$

where $\hat{F}_g$ is chosen as before. In a similar manner to the description above, this estimated bias can be used to adjust the estimate $\hat{F}_\infty$ to allow for the visible bias.

Tables 2.3 and 2.4 provide the simulated performance of these bias adjusted versions of the original estimators for the same simulations considered before. In constructing the bias adjusted estimators, a sample size of $n = 500$ was used in step (ii) of the algorithm described above and 1,000 simulations of this step were carried out. It is clear from the results reported in Tables 2.3 and 2.4 that the bias adjusted estimators have significantly improved the performance of both estimators in terms of bias, with only modest increases in variability. The improvement is more noticeable in Table 2.3 as the original bias is much greater. Note that the bias adjustments can also be calculated when $A = 0$ or $A = \infty$ but are not presented in these table. However, for completeness, the bias adjusted estimates are presented for these values of A in the additional applications in Appendix A.

2.2.5 Regression Models

We briefly consider an extension of the above ideas to the regression context where interest focuses on the effects of a (potentially multidimensional) covariate X. The use of regression models, parametric, semi-parametric and nonparametric, for current status data was described in Section 1.3 where applicable references were given, in which more details can be found. Much of the literature on current status data has exploited the correspondence between standard regression models for the underlying failure time and generalized linear models for the observed current status outcome. For ease of understanding, we use a specific parametric model to illustrate this association with generalized linear models. A similar approach can also be applied to other parametric models. Consider current status data with no misclassification and assume the following Weibull hazard function

$$h(t) = \exp(b)at^{a-1}.$$

Note thats

$$H(C) = \int_0^C h(t)dt = \exp(b)a \int_0^C t^{a-1}dt = \exp(b)C^a.$$

This can also be written in terms of the survival function where

$$S(C) = \exp(-H(C)) = \exp(-\exp(b)C^a).$$

Now assume a proportional hazards model such that

$$S(C) = [S_0(C)]^{e^{\lambda x}},$$

where S_0 is an arbitrary baseline survival function. We use λ here instead of β, the standard notation used to represent the regression coefficient, to avoid confusion with the classification rates used throughout this chapter. Combining these equations we get

$$[S_0(C)]^{e^{\lambda x}} = \exp(-\exp(b)C^a).$$

This equation can then be written as

$$\log(-\log(S_0(C)) = b + a\log(C) + \lambda x.$$

This represents a particular case of generalized linear model for Y with a complementary log-log link function. The advantage of this approach is that the regression coefficients λ are exactly the relative hazards from the regression model for T. Assuming a proportional odds regression model instead of a proportional hazards model would result in the use of the logit link function instead of the complementary log-log link function.

To adapt these techniques to incorporate constant misclassification we use the ideas of binary generalized linear models with outcomes subject to misclassification (Neuhaus, 1999). To understand how constant misclassification can be incorporated into the regression setting through modification of the link function, consider the following illustration where we assume proportional odds regression with a logit link function. The distribution function F and the logit link function can be defined in general as

$$F(c|X) = \frac{1}{1 + e^{-a(c)-b(x)}}, \qquad \text{logit link: } \log\left(\frac{p(x|c)}{1 - p(x|c)}\right),$$

where $p(x|c) = E[Y|C, X]$. In the absence of misclassification this implies

$$g\{P(Y = 1|X, C)\} = \log\left(\frac{\frac{1}{1+e^{-a(c)-b(x)}}}{1 - (\frac{1}{1+e^{-a(c)-b(x)}})}\right) = \log\left(\frac{1}{e^{-a(c)-b(x)}}\right) = a(c) + b(x).$$

Combining the definition of G (Section 2.2) with the proportional odds assumption gives

$$G(C|X) = (\alpha + \beta - 1)F(C|X) + (1 - \beta) = \left(\frac{\alpha + \beta - 1}{1 + e^{-a(c)-b(x)}}\right) + (1 - \beta).$$

Let $\gamma = \alpha + \beta - 1$ and define the modified link function as

$$\text{modified link function: } \log\left(\frac{(p(x|c) - (1-\beta))/\gamma}{1 - ((p(x|c) - (1-\beta))/\gamma)}\right).$$

This implies

$$\begin{aligned} g^*\{P(\Delta = 1|X, C)\} &= \log\left(\frac{((\frac{\gamma}{1+e^{-a(c)-b(x)}}) + (1-\beta) - (1-\beta))/\gamma}{1 - ((\frac{\gamma}{1+e^{-a(c)-b(x)}}) + (1-\beta) - (1-\beta))/\gamma}\right) \\ &= \log(e^{a(c)+b(x)}) = a(c) + b(x) \end{aligned}$$

More formally, if we focus on the constant misclassification model, and with the same assumptions as before, it follows that

$$P(\Delta = 1|X, C) = (\alpha + \beta - 1)P(Y = 1|X, C) + (1 - \beta)$$

and so

$$P(\Delta = 1|X, C) = (\alpha + \beta - 1)g^{-1}(\eta_{x,c}) + (1 - \beta),$$

where g is the link function in the induced generalized linear model for Y. In addition, in most models, the regression term $\eta_{x,c}$ is also additive in x and c (or $\log(c)$). It follows that the observed outcome Δ also follows a generalized linear model with a modified link function, namely

$$g^*(P(\Delta = 1|X, C)) = g\left\{\frac{P(\Delta = 1|X, C) - (1-\beta)}{(\alpha + \beta - 1)}\right\}.$$

For example, assuming a Weibull regression model for T, the generalized linear model for Y in X and C involves g, the complementary log-log link function. By making such parametric assumptions, the corresponding regression model will often simplify to a familiar generalized linear model. We fit regression models to the example on HPV infection status. In Model (a) we assume no errors in the response variable (therefore using g directly), and in Model (b) we adjust for constant misclassification with the known classification rates of $\alpha = 0.73$ and $\beta = 0.9$ (using g^*). These assumed classification rates allow for both laboratory error and the possibility that some negative tests fail to detect prior HPV infection, as discussed in Section 2.2.2. Note that the parameter estimates in both models have proportional hazards interpretations on age at first infection with HPV, according to the Weibull regression model assumption for T, as distinct from the simple cross-sectional interpretations discussed in Neuhaus (1999). The results of both models are presented in Table 2.5, along with the observed ratio of parameter estimates. The generalized linear model induced by Weibull regression indicates that age at screening must be included in the model additively on the log scale. The standard errors were obtained from the observed information matrix and were calculated using PROC GENMOD in SAS version 9.1.

Covariate	log (RH): Model (a) ($\hat{\beta}^*$) Ignoring misclassification	log(RH): Model (b) ($\hat{\beta}$) Adjusted for misclassification	$\hat{\beta}^*/\hat{\beta}$
Smoke now	0.056 (0.108)	0.103 (0.144)	0.544
STD	-0.479 (0.299)	-0.698(0.258)	0.686
log(age at screening)	0.822 (0.455)	1.269 (0.552)	0.648

Table 2.5: Estimates (and standard errors) of the log Relative Hazard (RH) for time to first HPV infection, which is assumed to follow a Weibull distribution. Model (a) ignores misclassification in the response variable and Model (b) incorporates constant misclassification corresponding to $\alpha = 0.73$ and $\beta = 0.9$.

According to Model (a) and Model (b), respectively, the hazard of first HPV infection is increased by 6% and 11% for those who currently smoke (Smoke now = 1) to those who do not smoke (Smoke now = 0), holding other covariates in the model fixed. Clearly this effect is not significant. On the other hand, the hazard of HPV infection is reduced by 38% and 50% for those who have had any other prior sexually transmitted disease (STD = 1) compared with those who have not (STD = 0). This effect is quite strikingly significant, at least when misclassification is accounted for. As reported by Neuhaus (1999), the ratio of the parameter estimates suggest that ignoring the errors in the HPV screening results leads to substantially biased estimates of the associations of covariates with infection status, with the direction of the bias reflecting attenuation towards the null. Our findings are qualitatively similar to those of Neuhaus (1999) although we show a somewhat lower effect for the presence of a prior STD, presumably due to our allowance for additional misclassification.

2.3 Discussion

We have discussed estimation of the NPMLE of a distribution function based on current status data where the outcome variable is subject to misclassification. The ideas were also easily extended to regression models for the underlying survival time. We have illustrated the latter using a parametric regression model. Alternative methods to allow for misclassification in the current status response include the simulation extrapolation (SIMEX) method (for the regression setting, see Hardin et al. (2003) where the SIMEX method is applied to standard generalized linear models). More recently, Küchenhoff et al. (2006) applied SIMEX to binary outcome data associated with a generalized linear model and compared results to the maximum likelihood approach espoused by Neuhaus (1999). Note also that the covariates could alternatively be observed with error. Wen et al. (2011) consider a semiparametric maximum likelihood method for analyzing current status data with mismeasured covariates under the proportional hazards assumption.

Although we considered a parametric regression model, semi-parametric survival models can also be analyzed using the ideas of Shiboski (1998a) on semi-parametric generalized additive models. In this case, the technique of adjusting the link function to allow for misclassification, discussed in Section 2.2.5, can also be used. SIMEX again provides an alternative approach. In addition, the bias adjusted algorithm discussed in Section 2.2.4 can also be applied in the regression context, in particular to allow for more more complex misclassification models. Sal y Rosas and Hughes (2011) extend the work of this chapter and consider nonparametric and semi-parametric analysis of current status data with the response variable subject to misclassification.

Throughout this chapter we have assumed that the misclassification rates and window of misclassification, if appropriate, are known exactly. In some cases, the rates may have to be estimated from a validation sample where the true response is measurable perhaps by use of an expensive 'gold standard' technique. This data can then be incorporated into a full likelihood that will then account for the uncertainty in estimation of the misclassification rates. In principal, a similar approach could be used for validation data that provided information on the value of A or the size of the misclassification window. However, estimation of the value of A is itself a much studied non-trivial estimation problem in detecting the time of transition of binomial classification rates. We leave these interesting extensions to future work.

Chapter 3

Misclassification of Current Status Data with Group Testing

3.1 Introduction

In many epidemiological applications individuals are screened for the presence of an infectious disease, or disease specific antibodies. In such studies, the available resources for testing are often limited, either by cost or other factors. With certain infectious diseases, including sexually transmitted diseases, infected individuals are often asymptomatic, which results in many individuals being left untreated and others becoming infected unknowingly. To address such problems, it is necessary to test a large number of individuals. Given the large number of tests and the associated cost of testing each individual sample, it is important to find ways to reduce the amount of testing needed without screening fewer individuals. Bilder et al. (2010) address this public health issue through the use of group testing and consider an application to the epidemic levels of chlamydia and gonorrhea in Nebraska (Zagurski, 2006). Group testing, also referred to as batch testing or pooled testing, was first introduced during the Second World War as a method for reducing the cost of detecting syphilis in U.S. soldiers (Dorfman, 1943).

Group testing combines samples, such as blood or urine, from a number of individuals and then tests the combined (group) sample rather than testing each individual sample for the presence of the disease of interest. The number of tests to be performed is therefore greatly reduced with this form of testing. The original group testing procedure of Dorfman (1943) involves initially testing the group sample, from which, if the group tests negative, then all individual samples within that group are declared negative. However, if the group tests positive, then each sample within that group is subsequently tested individually to identify the positive samples. Many variations of this strategy were later developed where different pooling algorithms and retesting mechanisms were applied to the positive sample, rather than retesting each individual sample (see for example, Brookmeyer (1999), Bilder et al. (2010), Xie et al. (2001), Hughes-Oliver and Swallow (1994), amongst others). The main advantage of group testing is that it can reduce costs when a large number of individuals need to be tested. Group testing also offers a feasible way to lower the error rates associated with certain testing mechanisms, particularly if the disease prevalence is low, for example when screening low-risk human immunodeficiency virus (HIV) populations (Litvak et al., 1994). Additional applications of group testing have been seen in blood bank screening, public health, genetics, drug discovery and development, and many other areas (Bilder et al., 2010). In the group testing framework, if interest is in estimating disease prevalence (the probability of a positive test result at a single point in time), when a positive group is identified, retesting individual samples within that group is not necessary for estimation of this prevalence. In this chapter we consider group testing from a current status data framework and also assume the retesting of positive groups is not performed.

In many applications appropriate for group testing, ascertainment of the current status outcome (the grouped result) is based on a screening test which may not have perfect sensitivity or specificity, as few diagnostic tests are perfect. It is therefore important to consider misclassification when using such forms of diagnostic testing. Chapter 2 describes misclassification of simple current status data, where samples are tested individually, using known misclassification rates. Misclassification of disease status in group testing has also received attention (Tu et al., 1995, Liu et al., 2011, Graff and Roeloffs, 1972), although it has not yet been specifically addressed from a current status data perspective. Section 3.2 shows how the current status data structure is directly applicable to the group testing strategy described in the literature.

In the remainder of this chapter, we develop on the previous work of other authors, relating to misclassification and group testing. The main focus of this chapter is in estimating the distribution function of time to a specific event, such as the time to disease infection, and specifically whether group testing can be used for more efficient estimation of this distribution function when the current status outcomes are subject to misclassification. In Section 3.2 we extend the notation of Chapter 2 to consider misclassification of current status data with group testing. Initially we consider the simplest group testing scenario where all individuals are tested at the same time, and all groups are equal in size (Section 3.3.1). We then extend this simple model to allow for the multiple monitoring time scenario, again with equal group sizes (Section 3.3.2). Determination of the optimal group size to be used under each of these scenarios is also discussed. In Section 3.3.3 we consider a more complex model where the group sizes are dependent on the monitoring times. We also briefly address the situation where the groups may be made up of individuals with different monitoring times. Finally, in Section 3.4 we address the current status regression setting, introduced in Section 1.3 and consider regression in the context of current status data with group testing and misclassification.

3.2 Notation and Likelihood

For current status data with a single event of interest, as described in Chapter 2, we assume the following notation. Let T be the survival time random variable of interest with the corresponding distribution function F, and let the monitoring time be denoted by the random variable C, which is assumed to be independent of T. In some examples C is non-random, but in either case, we generally focus on the conditional likelihood, given C. For the random monitoring time scenario, current status observation refers to a sampling scheme where n i.i.d. observations are collected on the random variable (Y^I, C) where $Y^I = I(T \leq C)$. This notation corresponds to each individual sample being tested separately. The binary variable Y^I is therefore observed for each of the n individual samples. This individual level response variable Y^I directly corresponds to the binary variable Y of Chapter 2.

To implement the group testing strategy of Section 3.1 we use the following notation. Again, define the monitoring time as C, the time at which individuals are tested. Define the sample size as n, the total number of individuals sampled, or the total number of specimens obtained. Suppose these individuals are divided into m groups where each group has the corresponding group size k_j, for $j = 1, 2, ..., m$. Therefore, the total sample size can be defined as $n = \sum_{j=1}^{m} k_j$. This allows for the possibility of differing group sizes, which may in fact depend on the monitoring times. This idea is addressed further in Section 3.3.3. In many examples, both in the literature and in practice, it is assumed that all groups are of equal size, say size k, in which case the definition of the sample size can be simplified to $n = mk$. In the group testing scenario, instead of observing the outcome Y^I for each individual we now observe a binary variable Y for each of the m batches or groups. As described above, with the group testing approach if the outcome for group j is negative ($Y_j = 0$), then all samples within that group are defined as having a negative result. This means if $Y_j = 0$, then $Y^I = 0$ for all individuals in group j. However, recall that a positive group test result does not imply a positive test result for all individuals in that group, it simply indicates there is at least one positive individual within the group. Therefore, if the group outcome $Y_j = 1$ then $Y^I = 1$ for at least one of the k_j individuals in group j. Without further testing, identification of the positive individual(s) is not possible.

Motivated by the need for group testing discussed in Section 3.1, and the potential for misclassification in diagnostic tests, we now consider the possibility that the outcome variable is observed with error, applicable to both the individual sample and group testing scenarios. We focus primarily on the constant misclassification model, similar to that described in Section 2.2, where we assume the classification rates are known and fixed. These methods can also be extended to allow for alternative error models, such as the time-varying misclassification model considered in Section 2.2.3. To allow for misclassification, assume that instead of observing the correct response Y^I or Y we now observe the error corrupted response Δ^I or Δ, respectively. We assume the classification rates are inherent characteristics of the testing mechanism and that the sensitivity and specificity rates are not affected by grouping. This means that the classification rates remain fixed regardless of the group size. This may not be true for all testing mechanisms, and in some cases there may be a maximum group size up to which these classification rates hold. For example, Kline et al. (1989) showed the sensitivity and specificity of the ELISA, a test used for the detection of HIV, is not affected by pooling samples up to a size of $k = 15$, for a specific test kit and level of dilution. For group sizes greater than a defined maximum value, higher levels of misclassification may be observed. In Section 3.3.1 we briefly discuss this case where the classification rates may increase as the group size increases.

First, consider constant misclassification where increasing the group size does not alter the accuracy of the assumed classification rates. These classification rates can be defined as follows, for both the individual and grouped samples,

$$\begin{aligned} P(\Delta^I = 1|Y^I = 1) &= P(\Delta = 1|Y = 1) = \alpha \\ P(\Delta^I = 0|Y^I = 0) &= P(\Delta = 0|Y = 0) = \beta, \end{aligned}$$

where Y^I and Y are the true individual and group outcomes, with Δ^I and Δ representing the observed response for the individual and group samples, respectively. With this type of group testing, the observed data can be defined as m i.i.d. copies of (Δ, C), where $1 \leq m \leq n$. If there exists equal group sizes with $k_j = 1$ for all j, then $Y^I = Y$ and $\Delta^I = \Delta$, resulting in this setup reducing directly to the misclassified current status data setting discussed in Chapter 2.

Following the notation and assumptions of Chapter 2, assume the classification probabilities α, $\beta > 0.5$ and do not depend on the monitoring time, or the number of individuals within a group. Assume also that only individuals with the same monitoring time are grouped together, so the entire group can be indexed by a single monitoring time. Therefore, if C_m is the monitoring time for group m, then all individuals within this group are observed at time C_m. We discuss an extension of this assumption in Section 3.3.4. Let C_j be the j^{th} order statistic of $C_1, C_2, ..., C_m$ and let δ_j be the observed value of Δ for group j. The (conditional) likelihood function for this misclassification structure is given by

$$\prod_{j=1}^{m} [P(\Delta_j = 1|c_j)]^{\delta_j} [P(\Delta_j = 0|c_j)]^{1-\delta_j}, \tag{3.1}$$

where c_j is the observed value of C_j, and m is the total number of groups examined. The probabilities $P(\Delta_j = 1|c_j)$ and $P(\Delta_j = 0|c_j)$ can be defined as follows

$$\begin{aligned} P(\Delta_j = 1|c_j) &= P(\Delta_j = 1|y_j = 1, c_j)P(y_j = 1|c_j) + P(\Delta_j = 1|y_j = 0, c_j)P(y_j = 0|c_j) \\ &= \alpha - (\alpha + \beta - 1)(1 - F(c_j))^{k_j}, \end{aligned}$$

and

$$\begin{aligned} P(\Delta_j = 0|c_j) &= P(\Delta_j = 0|y_j = 0, c_j)P(y_j = 0|c_j) + P(\Delta_j = 0|y_j = 1, c_j)P(y_j = 1|c_j) \\ &= 1 - \alpha + (\alpha + \beta - 1)(1 - F(c_j))^{k_j}. \end{aligned}$$

Note that by setting $k_j = 1$ for all j, these probabilities and the corresponding likelihood given in Equation (3.1) will be equal to the probabilities and likelihood of Equation (2.1). For ease of notation, let $\gamma = \alpha + \beta - 1 > 0$. The (conditional) likelihood function allowing for constant misclassification in the response variable of a grouped sample is then

$$\prod_{j=1}^{m} [\alpha - \gamma(1 - F(c_j))^{k_j}]^{\delta_j} [1 - \alpha + \gamma(1 - F(c_j))^{k_j}]^{1-\delta_j},$$

with the corresponding log-likelihood

$$\sum_{j=1}^{m} \delta_j \log(\alpha - \gamma(1 - F(c_j))^{k_j}) + \sum_{j=1}^{m} (1 - \delta_j) \log(1 - \alpha + \gamma(1 - F(c_j))^{k_j}). \qquad (3.2)$$

In the following section we consider maximization of this log-likelihood under various scenarios. Section 3.3.1 considers the group testing scenario where there is a single monitoring time and all groups contain the same number of individuals. This setting is extended in Section 3.3.2 to allow for multiple monitoring times, again assuming all groups are equal in size. Section 3.3.3 relaxes the assumption of equal group sizes and allows the group size to vary with the monitoring times. Finally, Section 3.3.4 discusses maximization of this likelihood (Equation (3.2)) under the more complex scenario where individuals with different monitoring times can be tested together as one group.

3.3 Nonparametric Estimation of a Single Distribution Function

As mentioned above, there are various group testing approaches that can be considered. In this section we consider various group testing scenarios and show how the complexity of maximizing the above log-likelihood (Equation (3.2)) depends on whether there are single or multiple monitoring times, and whether the group sizes are equal or vary with the monitoring time. In each case, we assume the classification rates are known and constant and do not depend on the number of individuals within a group, though the concept of misclassification increasing with sample size is briefly discussed in Section 3.3.1. Throughout this chapter, our focus is primarily on nonparametric estimation of the distribution function F. This chapter is therefore an extension of Chapter 2 and many of the ideas previously used can therefore be extended to this group testing scenario. We also compare the results of this group testing approach to that of testing each sample individually (Chapter 2), from which we can assess the situations in which it is beneficial to use group testing as an alternative testing approach.

To obtain the nonparametric maximum likelihood estimate (NPMLE) of F, we seek to maximize the log-likelihood in Equation (3.2). We follow the general approach of Section

2.2, previously used for simple current status data subject to misclassification. In this group testing scenario, define the distribution function G as

$$G(c_j) \equiv \alpha - \gamma(1 - F(c_j))^{k_j},$$

where $\gamma = \alpha + \beta - 1$. Consequently, as before, the NPMLE of the distribution function F, when the current status outcomes are subject to misclassification, can be found by obtaining a vector $\tilde{z} = (z_1 = \hat{G}(c_1), \ldots, z_m = \hat{G}(c_m)) \in R^m$ maximizing

$$\phi(G(c_j)) = \sum_{j=1}^{m} \delta_j \log(G(c_j)) + \sum_{j=1}^{m} (1 - \delta_j) \log(1 - G(c_j)) \tag{3.3}$$

under the constraint

$$1 - \beta \leq G(c_1) \leq G(c_2) \leq \cdots \leq G(c_m) \leq \alpha. \tag{3.4}$$

Note that Equation (3.3) is similar to Equation (2.2), with the indices now differing. Equation (3.4) is identical to Equation (2.3), though the definition of the distribution function G in the group testing scenario is now complicated by the additional term defining the group size. However, even with this alternative definition of G, it is easily seen that the constraints required in maximizing the log-likelihood of Equation (3.3) remain unchanged, regardless of the group size. Note that the constraints are designed to control the estimates close to the extremes of the distribution. In the limit, as $F \to 0$ or $F \to 1$, individuals will either all be negative ($F = 0$) or will all be positive ($F = 1$). Therefore, at these values, the grouping mechanism will not alter the estimation.

3.3.1 Single Monitoring Time, Single Group Size

Consider the simplest group testing scenario where there is a single monitoring time C for all individuals, and a single group size for all combined samples. Let k denote the number of individuals in each group and assume the total sample size n is made up of exactly m groups of size k. We therefore assume $m = n/k$ is an integer. This assumption is also made by Tu et al. (1995) and Liu et al. (2011) in estimating the prevalence of a rare disease. For this simple group testing scenario, the definition of G reduces to $G(C) \equiv \alpha - \gamma(1 - F(C))^k$. A claim, and corresponding proof, similar to that of Section 2.2 can be applied and followed, where the resulting estimate of the NPMLE of F at the monitoring time C is then obtained through the use of the following relationship

$$\widehat{F}(C) = 1 - \left(\frac{\alpha - \widehat{G}(C)}{\alpha + \beta - 1} \right)^{1/k}, \tag{3.5}$$

where, as in Chapter 2, G can be estimated using the straightforward PAV algorithm (see Section 1.2), with all estimates of G falling between $1-\beta$ and α. Estimates not falling within this region are shifted to equal the appropriate bound.

As stated by Liu et al. (2011), in the presence of perfect classification, group testing can substantially reduce the cost incurred with testing a large number of individuals, but will always yield a less efficient estimate of prevalence than that obtained by using fully observed data in which disease status is determined for each individual. As mentioned in Section 3.1, imperfect sensitivity or specificity are often inherent in many epidemiological applications. Tu et al. (1995) show that in the presence of misclassification of disease status, group testing can be used for more efficient estimation of the prevalence of a rare disease, as well as reducing the number of necessary tests. Tu et al. (1995), Litvak et al. (1994) and Liu et al. (2011) all make assumptions equivalent to assuming the samples are pooled into groups of equal size (k), the number of groups ($m = n/k$) is an integer, and all observations are obtained at a single monitoring time (C). Note that current status data with a single monitoring time leads directly to estimation of prevalence at this monitoring time. Therefore, although the notation and approach differ slightly, it is not surprising that the estimate of rare disease prevalence defined by Tu et al. (1995) is equivalent to our estimate of $\widehat{F}(C)$ given in Equation (3.5).

Liu et al. (2011) extend the results of Tu et al. (1995) to determine the level of prevalence, for groups of equal sizes, below which group testing is more efficient than (a) random sampling and (b) the fully observed data. We now add to these results by determining the maximum group size, below which group testing is preferred to testing each individual separately. We also determine the optimal group size for estimation of the distribution function at a single point in time, using known classification rates. For comparison with the work of Tu et al. (1995) and Liu et al. (2011) we assume the total n samples are divided into m groups of size k, all of whom are examined at time C. We assume there are no further limitations on the possible group sizes. To validate this assumption, the testing mechanism may need to be examined to determine whether increasing the pool size alters the sensitivity and specificity of the test. As mentioned previously, in an application to HIV screening, Kline et al. (1989) show, for a specific test and level of dilution, that the sensitivity is not affected by pooling samples up to a size of $k = 15$. Many authors therefore limit the possible group sizes to a maximum of $k = 15$, as the accuracy beyond this level of grouping may not be equal to that of a single sample test. If no such restriction is imposed by the testing mechanism, the group size k can range from $k = 1$ to $k = n$.

In our determination of this maximum group size, and the optimal group size for estimation of the distribution function at a single monitoring time, we compare the mean squared error (MSE) obtained for the various group sizes to that of the single sample analysis ($k = 1$). The calculated maximum group size is therefore the largest group size with a MSE lower than the MSE found when $k = 1$. It is important to note that all group sizes less than this maximum will also produce more efficient estimates, so all smaller group sizes are at least as good as the single sample setting. The optimal group size is the group size with the lowest MSE amongst all possible group sizes.

Table 3.1 presents the maximum group size (optimal group size) for estimation of a distribution function at a single point in time, or equivalently, the prevalence at a single point in time. The results presented are based on simulated data at three unique monitoring times, where the corresponding values of F at these monitoring times are $F(C) = \{0.01, 0.04, 0.1\}$. For each of these monitoring times both symmetric and asymmetric classification rates are considered. For the situation of asymmetric classification, both cases, $\alpha > \beta$ and $\alpha < \beta$, are examined, as both scenarios are plausible in practice and depend on the specific application. As can be seen in Table 3.1, we also consider how these group sizes change as the sample size changes. We examine a variety of sample sizes from $n = 200$ to $n = 2000$. For each of these sample sizes, all possible integers, $m = n/k$, are considered, including $k = 1$ which corresponds to the simple current status data setting of Chapter 2. It must be noted that the number of possible group sizes vary depending on the total sample size and so certain group sizes are not considered in all cases. The presented results are based on the MSE obtained through simulated data where all simulations are repeated 1000 times. It should be noted that in these simulations, the grouping is assumed to occur prior to adjusting the observations for misclassification. Therefore, it is the grouped outcomes that are subject to the pre-defined levels of misclassification.

These results compare, in the presence of misclassification, the use of group testing with the standard approach of testing each sample individually, as described in Chapter 2. It can be seen from Table 3.1 that both the maximum and optimal group sizes depend on the value of the distribution function being estimated, the total sample size and the levels of sensitivity and specificity. These results show that as the sample size increases, both the maximum and optimal group sizes also increase (or at least are non-decreasing), regardless of the levels of classification, α and β. However, the magnitude, and amount by which these group sizes increase is influenced by the α and β classification rates. It is also seen that as the value of F increases, the maximum (optimal) group size decreases, keeping the classification rates fixed. The greatest gain therefore appears to occur when the monitoring time is small. Within each monitoring time, the highest group sizes observed occur when the sensitivity is greater than the specificity, namely when $\alpha = 0.95$, $\beta = 0.8$.

n	200	500	700	1200	1500	2000
$F(C) = 0.010$						
$\alpha = 0.95$, $\beta = 0.80$	25 (25)	50 (50)	70 (70)	100 (100)	125 (125)	125 (125)
$\alpha = 0.80$, $\beta = 0.95$	25 (20)	50 (25)	50 (28)	60 (24)	60 (50)	50 (50)
$\alpha = 0.95$, $\beta = 0.95$	25 (8)	50 (10)	50 (20)	80 (20)	75 (25)	100 (25)
$\alpha = 0.80$, $\beta = 0.80$	10 (10)	25 (20)	35 (28)	50 (24)	50 (30)	50 (50)
$F(C) = 0.039$						
$\alpha = 0.95$, $\beta = 0.80$	20 (20)	25 (25)	35 (35)	50 (30)	50 (30)	50 (40)
$\alpha = 0.80$, $\beta = 0.95$	10 (8)	20 (10)	20 (14)	25 (15)	25 (15)	25 (16)
$\alpha = 0.95$, $\beta = 0.95$	10 (4)	20 (4)	20 (5)	25 (6)	25 (6)	25 (5)
$\alpha = 0.80$, $\beta = 0.80$	8 (4)	10 (10)	20 (10)	20 (10)	25 (12)	25 (10)
$F(C) = 0.100$						
$\alpha = 0.95$, $\beta = 0.80$	10 (10)	20 (10)	25 (14)	30 (12)	30 (12)	40 (10)
$\alpha = 0.80$, $\beta = 0.95$	5 (4)	5 (5)	10 (7)	10 (6)	10 (6)	10 (8)
$\alpha = 0.95$, $\beta = 0.95$	4 (3)	5 (4)	5 (4)	5 (2)	3 (3)	5 (4)
$\alpha = 0.80$, $\beta = 0.80$	5 (4)	5 (4)	7 (5)	8 (4)	10 (5)	10 (5)

Table 3.1: Results from simulated data showing the maximum group size (optimal group size) at which group testing is preferred to examining each individual sample for the purpose of estimating the distribution function F at a single monitoring time. Three monitoring times are considered where the corresponding values of $F(C) = \{0.010, 0.039, 0.100\}$. Various sample sizes are considered from $n = 200$ to $n = 2000$. The results are given for a variety of known classification rates, considering both symmetric and asymmetric misclassification.

There are a few instances in Table 3.1 in which the results appear to decrease as the sample size increases. However, it must be noted that we only consider group sizes k such that $m = n/k$ is an integer. Therefore, the potential group sizes vary according to sample size. For example, with $F(C) = 0.010$, $\alpha = 0.80$ and $\beta = 0.95$, a maximum group size of $k = 60$ is calculated for the sample size $n = 1500$. This maximum decreases to $k = 50$ for the sample size of $n = 2000$. With this sample size, a group size of $k = 60$ violates the assumption that the number of groups $m = n/k$ is an integer. The next possible group size would be $k = 80$, therefore the maximum group size presumably lies between $k = 50$ and $k = 80$. An extension of these results, not yet considered, is whether a mixture of group sizes at a single monitoring time could improve the estimate of F. For example, with $n = 2000$, 40 groups of size 50 could be compared with 33 groups of size 60 plus one group of size 20, to determine the optimal combination of group sizes for estimation of F at time C. Table 3.1 shows the maximum (optimal) group sizes when all groups are of equal size only.

The results presented in Table 3.1 assume the classification rates are not altered by increasing the group size. However, in practice this assumption may not always be accurate. As previously mentioned, in certain cases there may be a limiting group size above which there is greater potential for misclassification. Suppose the maximum group size for which the classification rates remain unchanged is defined at $k = 15$ (Kline et al., 1989). Note that many of the maximum (optimal) group sizes presented in Table 3.1 are greater than $k = 15$. This suggests that in certain cases, even if higher rates of misclassification are observed for larger group sizes, the optimal choice of group size may in fact be greater than $k = 15$. The estimates of the MSE used to determine the values of the maximum and optimal group sizes presented in Table 3.1 can be used to determine whether larger group sizes are preferable, even when higher rates of misclassification are induced.

3.3.2 Multiple Monitoring Times, Single Group Size

We now extend this simple group testing scenario, where there exists a single monitoring time C and a single group size k, to consider the multiple monitoring time scenario. Under this scenario we are interested in estimating the underlying distribution function F, and not simply the prevalence at a single point in time. In this section, although we assume multiple monitoring times exist, we retain the assumption of equal group sizes k, and again assume all individuals within a specific group have the same monitoring time. Under this scenario we define the distribution function G as

$$G(c_j) \equiv \alpha - \gamma(1 - F(c_j))^k,$$

where c_j is the monitoring time for the jth group. Under this definition of G, we maximize the likelihood of Equation (3.3) by estimating G using the PAV algorithm with the constraints defined in Equation (3.4). The resulting estimate of F at c_j is then given by

$$\widehat{F}(c_j) = 1 - \left(\frac{\alpha - \widehat{G}(c_j)}{\alpha + \beta - 1} \right)^{1/k}. \tag{3.6}$$

This approach therefore allows the distribution function to be estimated at more than a single point in time. As noted by Tu et al. (1995), and seen in Section 3.3.1, group testing can be used to produce more efficient estimates of the prevalence of a rare disease. We therefore also consider the rare disease setting in this multiple monitoring time scenario. With a rare disease, we expect the values of the true underlying distribution function to have low values at many of the monitoring times since $F = P(T \leq C)$. To determine whether group testing is a useful tool for estimation of the distribution function we examine the results of simulated data, presented in Table 3.2 and Table 3.3. Interest could alternatively be in estimating a functional of F, for which similar simulations could be performed. We simulate hypothetical data where we assume the true underlying distribution follows an Exponential distribution with mean 100 to represent a distribution function describing a rare disease setting.

C	5	10	15	20	25
F(C)	0.049	0.095	0.139	0.181	0.221
$\alpha = 0.95,\ \beta = 0.90$					
$k = 1$	0.005 (0.009)	0.035 (0.032)	0.080 (0.032)	0.121 (0.032)	0.163 (0.032)
$k = 2$	0.020 (0.017)	0.063 (0.032)	0.104 (0.032)	0.145 (0.032)	0.188 (0.032)
$k = 4$	0.032 (0.017)	0.076 (0.032)	0.117 (0.032)	0.157 (0.032)	0.198 (0.032)
$k = 5$	0.034 (0.017)	0.078 (0.032)	0.120 (0.032)	0.159 (0.032)	0.198 (0.032)
$k = 8$	0.039 (0.018)	0.083 (0.032)	0.123 (0.032)	0.159 (0.032)	0.215 (0.118)
$k = 10$	0.040 (0.018)	0.083 (0.032)	0.121 (0.032)	0.159 (0.063)	0.281 (0.262)
$k = 20$	0.045 (0.021)	0.095 (0.122)	0.174 (0.243)	0.349 (0.063)	0.521 (0.437)
$k = 25$	0.049 (0.071)	0.125 (0.210)	0.220 (0.318)	0.334 (0.392)	0.537 (0.447)
$\alpha = 0.90,\ \beta = 0.95$					
$k = 1$	0.107 (0.032)	0.154 (0.032)	0.197 (0.032)	0.240 (0.032)	0.283 (0.032)
$k = 2$	0.079 (0.032)	0.128 (0.032)	0.173 (0.032)	0.217 (0.032)	0.263 (0.032)
$k = 4$	0.067 (0.022)	0.117 (0.032)	0.166 (0.032)	0.211 (0.032)	0.265 (0.045)
$k = 5$	0.064 (0.021)	0.115 (0.032)	0.163 (0.032)	0.212 (0.032)	0.274 (0.084)
$k = 8$	0.062 (0.021)	0.112 (0.032)	0.171 (0.045)	0.248 (0.138)	0.435 (0.091)
$k = 10$	0.060 (0.022)	0.114 (0.032)	0.191 (0.138)	0.375 (0.324)	0.685 (0.142)
$k = 20$	0.082 (0.154)	0.405 (0.430)	0.730 (0.407)	0.901 (0.277)	0.969 (0.301)
$k = 25$	0.115 (0.188)	0.403 (0.423)	0.681 (0.424)	0.843 (0.336)	0.933 (0.232)
$\alpha = 0.90,\ \beta = 1$					
$k = 1$	0.157 (0.032)	0.204 (0.032)	0.250 (0.032)	0.291 (0.032)	0.336 (0.032)
$k = 2$	0.109 (0.024)	0.159 (0.032)	0.206 (0.032)	0.252 (0.032)	0.301 (0.032)
$k = 4$	0.084 (0.022)	0.135 (0.032)	0.188 (0.032)	0.239 (0.032)	0.298 (0.055)
$k = 5$	0.078 (0.022)	0.130 (0.032)	0.187 (0.032)	0.242 (0.045)	0.320 (0.100)
$k = 8$	0.071 (0.023)	0.130 (0.032)	0.201 (0.077)	0.338 (0.230)	0.616 (0.342)
$k = 10$	0.069 (0.032)	0.136 (0.071)	0.303 (0.277)	0.618 (0.382)	0.882 (0.270)
$k = 20$	0.123 (0.226)	0.677 (0.432)	0.968 (0.163)	1.000 (0.021)	1.000 (0.022)
$k = 25$	0.144 (0.270)	0.595 (0.439)	0.936 (0.227)	1.000 (0.021)	1.000 (0.022)

Table 3.2: Simulation averages (standard deviations) of the estimates of the distribution function F (Exponential with mean 100) at 5 monitoring times ($C = \{5, 10, 15, 20, 25\}$), when the data is subject to misclassification and group testing. The assumed sample size is $n = 1000$ with 200 individuals observed at each monitoring time. A variety of classification rates are considered. For each defined level of misclassification, results are presented for group sizes $k = \{1, 2, 4, 5, 8, 10, 20, 25\}$.

The data used in Table 3.2 was generated as follows. Let T be Exponentially distributed, with mean 100 and assume a total sample size of $n = 1000$, with five unique monitoring times of interest ($C = \{5, 10, 15, 20, 25\}$). Assume each monitoring time observes an equal, fixed number of 200 individuals. As noted Section 3.2, the likelihood we seek to maximize (Equation (3.3)) assumes the grouped samples are made up of individuals all sharing the same monitoring time. Therefore, at each monitoring time there exists $m = 200/k$ grouped samples, again where k is the group size, and m is an integer. For each pair of T and C values, the corresponding (correct) outcome Y is produced, where $Y = I(T \leq C)$. At each monitoring time, the Y values are grouped into m batches of size k. The group outcome is then defined as positive if at least one of the individual samples within the group shows a positive result. The group results are then misclassified according to the pre-defined α and β classification rates. We assume these classification rates hold and are fixed for all group sizes. The estimate of F based on this form of group testing is then obtained through the relationship defined in Equation (3.6), with the appropriate estimate of G obtained through the PAV algorithm.

All possible group sizes, $k = \{1, 2, 4, 5, 8, 10, 20, 25, 40, 50, 100, 200\}$, were analyzed. However, only a selection are presented in Table 3.2. Group sizes above $k = 25$ are not shown as it is evident that, even at very low monitoring times, such large amounts of grouping are not beneficial. This is as expected, as for example, at $k = 200$ this would imply all individuals are grouped into a single sample, thereby limiting the information at a specific monitoring time to a single binary outcome. A variety of classification rates are considered and presented in Table 3.2 for each of the assumed group sizes. The simulation averages and standard deviations of the estimates of the distribution function F at the five monitoring times are given in the table. The true value of F at each of these monitoring time is also given for comparison. For each monitoring time and classification rate combination, the group size with the lowest MSE is highlighted by an underline.

Table 3.3 gives the results of similar simulations to those described above, where T is assumed Exponentially distributed with mean 100, $n = 1000$, and five monitoring times are considered with 200 individuals observed at each of these monitoring times. The differing aspect of this table is that later monitoring times are presented. Here, the assumed values of the monitoring times are $C = \{10, 20, 30, 40, 50\}$. The corresponding value of F at each of the assumed monitoring times is presented in this table along with the simulation averages and standard deviations for each combination of group size and classification rates. By examining the underlined values (those with the lowest MSE), Table 3.2 suggests that at early monitoring times, where the corresponding value of the underlying distribution function is low, group testing may be preferred to testing each sample individually ($k = 1$). Table 3.3 indicates that when group testing is used at later monitoring times, the individual estimates of F at these monitoring times will contain more bias than the estimates at the earlier monitoring times. Therefore, if interest is in estimating the distribution function of a non-rare disease, it is likely the only benefit to be obtained from the use of group testing is the reduction in cost and the number of tests to be performed since for non-rare diseases the true value of F at all monitoring times may be relatively high.

C	10	20	30	40	50
F(C)	0.095	0.181	0.259	0.330	0.393
$\alpha = 0.95,\ \beta = 0.90$					
$k = 1$	0.037 (0.032)	0.123 (0.032)	0.200 (0.032)	0.269 (0.032)	0.338 (0.045)
$k = 2$	0.063 (0.032)	0.146 (0.032)	0.221 (0.032)	0.289 (0.032)	0.350 (0.045)
$k = 4$	0.075 (0.032)	0.155 (0.023)	0.225 (0.032)	0.287 (0.045)	0.350 (0.055)
$k = 5$	0.080 (0.032)	0.159 (0.032)	0.225 (0.045)	0.288 (0.045)	0.358 (0.114)
$k = 8$	0.083 (0.032)	0.157 (0.032)	0.217 (0.071)	0.288 (0.155)	0.430 (0.290)
$k = 10$	0.086 (0.032)	0.156 (0.063)	0.224 (0.152)	0.329 (0.274)	0.552 (0.379)
$k = 20$	0.099 (0.122)	0.215 (0.288)	0.286 (0.243)	0.401 (0.406)	0.551 (0.435)
$k = 25$	0.118 (0.200)	0.202 (0.295)	0.281 (0.359)	0.385 (0.411)	0.568 (0.444)
$\alpha = 0.90,\ \beta = 0.95$					
$k = 1$	0.153 (0.032)	0.241 (0.032)	0.317 (0.032)	0.389 (0.032)	0.455 (0.032)
$k = 2$	0.128 (0.032)	0.219 (0.032)	0.303 (0.032)	0.377 (0.045)	0.449 (0.045)
$k = 4$	0.117 (0.032)	0.211 (0.045)	0.301 (0.045)	0.397 (0.077)	0.556 (0.200)
$k = 5$	0.115 (0.032)	0.212 (0.045)	0.316 (0.077)	0.466 (0.204)	0.727 (0.277)
$k = 8$	0.117 (0.032)	0.248 (0.141)	0.526 (0.330)	0.805 (0.303)	0.950 (0.170)
$k = 10$	0.120 (0.055)	0.413 (0.351)	0.769 (0.348)	0.921 (0.225)	0.978 (0.122)
$k = 20$	0.454 (0.448)	0.865 (0.317)	0.945 (0.212)	0.973 (0.148)	0.987 (0.105)
$k = 25$	0.403 (0.423)	0.720 (0.410)	0.855 (0.326)	0.929 (0.239)	0.963 (0.176)
$\alpha = 0.90,\ \beta = 1$					
$k = 1$	0.206 (0.032)	0292 (0.032)	0.371 (0.032)	0.441 (0.032)	0.504 (0.032)
$k = 2$	0.160 (0.032)	0.254 (0.032)	0.338 (0.045)	0.419 (0.045)	0.495 (0.055)
$k = 4$	0.137 (0.032)	0.239 (0.045)	0.345 (0.055)	0.484 (0.134)	0.731 (0.232)
$k = 5$	0.134 (0.032)	0.244 (0.045)	0.386 (0.138)	0.672 (0.272)	0.930 (0.173)
$k = 8$	0.134 (0.039)	0.356 (0.257)	0.791 (0.308)	0.982 (0.105)	0.999 (0.017)
$k = 10$	0.146 (0.109)	0.652 (0.389)	0.964 (0.158)	0.999 (0.021)	1.000 (0.021)
$k = 20$	0.703 (0.427)	0.995 (0.071)	1.000 (0.021)	1.000 (0.022)	1.000 (0.017)
$k = 25$	0.594 (0.439)	0.995 (0.071)	0.999 (0.032)	1.000 (0.022)	1.000 (0.022)

Table 3.3: Simulation averages (standard deviations) of the estimates of the distribution function F (Exponential with mean 100) at 5 monitoring times ($C = \{10, 20, 30, 40, 50\}$), when the data is subject to misclassification and group testing. The assumed sample size is $n = 1000$ with 200 individuals observed at each monitoring time. A variety of classification rates are considered. For each defined level of misclassification, results are presented for group sizes $k = \{1, 2, 4, 5, 8, 10, 20, 25\}$.

For the simple group testing scenario of Section 3.3.1, where we consider estimation of F at a single monitoring time, the optimal choice of group size k can be determined by direct comparison of the mean squared errors obtained for the different group sizes under consideration. However, in this case with multiple monitoring times, determining the optimal choice of k is not so straightforward. This optimal k depends on the question of interest, whether we are interested in estimating the distribution function at a single point in time, the entire distribution function F, or a functional of F, such as the mean. The nature of the question of interest will impact the optimal choice of k and ultimately whether the use of group testing is beneficial for reasons other than reduction in cost. Examining the individual point estimates in Table 3.2 and Table 3.3 we can see how this optimal group size (lowest MSE) is not consistent across all monitoring times. This optimal group size will therefore be determined by the functional of interest. For example, if we are interested in the conditional mean of F, conditional on the monitoring time being less than a specified value, say $C = 10$, Table 3.2 suggests group testing would be preferred to testing individual samples separately. However, if the mean of the entire distribution of F is of interest, although group testing is preferred at the early monitoring times, it is unclear whether the use of group testing would be beneficial.

In principle, a bias adjusted algorithm, such as that described in Section 2.2.3, could be used to remove the bias in Table 3.2 and Table 3.3. It should also be noted that these results are specific to an underlying Exponential distribution and therefore only represent the case of a constant hazard. Further simulations are necessary to examine how group testing would perform under an increasing or decreasing hazard.

3.3.3 Multiple Monitoring Times, Multiple Group Sizes

Previously, in Section 3.3.1 and Section 3.3.2, we assume a common group size k. However, due to many factors this assumption may be difficult to hold true in practice, and of course it may not be feasible to test equal group sizes at each monitoring time, especially under the assumption that $m = n/k$ is an integer. In this section we consider the more complex group testing scenario where the group size is determined by the monitoring time. It is therefore possible to have a different group size at each unique monitoring time. Further motivation for considering this group testing approach is such that, if determination of the group size is within the control of the examiner, varying the group size according to the monitoring time may in fact produce more efficient estimates of the underlying distribution function. Consideration should also be given to the appropriate choice of group size at each monitoring time when designing a study, particularly if resources are limited. The results of the previous tables suggest that larger group sizes should be used at the earlier monitoring times. Determining the optimal number of group tests to perform at each monitoring time could greatly reduce costs without a loss of accuracy in the estimates obtained.

Under this group testing scenario, let k_j denote the group size used at monitoring time c_j. To maximize the log-likelihood of Equation (3.3) under this group testing scenario, define the distribution function G as

$$G(c_j) \equiv \alpha - \gamma(1 - F(c_j))^{k_j}.$$

When multiple monitoring times exist along with multiple group sizes, although the constraints given in Equation (3.4) remain unchanged, for reasons described in Section 3.3, maximization of Equation (3.3) is now complicated by the fact that the group sizes k_j are not necessarily monotonic. Due to the varying group sizes, even if G is estimated as monotonic using a standard algorithm, the corresponding estimate of F may not be monotonic if we assume a simple direct relationship between G and F, as seen in the previous sections. When a single group size exists, G can be estimated using the straightforward PAV algorithm (see Section 1.2), with the corresponding estimate of F then easily obtained through Equation (3.5) or Equation (3.6) as the group sizes are all equal and therefore do not cause any monotonicity constraint violations. In this more complex setting, a direct estimate of F is not obvious to us.

As a towards estimating F, note that in the standard current status scenario with no misclassification, by assuming a proportional hazards model with no covariates, we can obtain an estimate of the distribution function F. This approach results in estimating a monotonic distribution function, for which the PAV algorithm can be applied. Introducing misclassification also allows the use of the PAV algorithm but with modified constraints that now depend on the classification rates α and β. With the introduction of group testing, ignoring misclassification temporarily, we examine estimating the distribution function F for group testing with group sizes depending on the monitoring times. Consider a proportional hazards model in the absence of covariates. For this group testing approach with no misclassification, define the distribution function G at time c_j as

$$G(c_j) = (1 - F(c_j))^{k_j}.$$

Applying the complementary log-log without making any assumptions produces the following

$$\begin{aligned} \log(G(c_j)) &= k_j \log(1 - F(c_j)) \\ \log(-\log(G(c_j)) &= \log(-k_j \log(1 - F(c_j)) = \log(-\log(1 - F(c_j)) + \log(k_j) \end{aligned}$$

Consequently, our estimation problem follows the general form of

$$g\{E[Y = 1|k, C]\} = H(C) + \log(k),$$

where g is a known link function and $H(C)$ is monotonic, from which we wish to estimate $H(C)$. Therefore, without misclassification, the problem of estimating F becomes a monotonic binary regression problem with a known offset ($\log(k)$) that differs depending on the monitoring time. The solution to such a problem is currently unknown to us, it would be of great interest to determine such an algorithm to further explore this group testing scenario.

With the introduction of constant misclassification, as described in Section 3.2, we can consider a similar analysis where the group size k_j separates additively, on the log scale. As seen in Section 2.2.5, misclassification can be incorporated through a simple modification of the link function. It follows that this form of group testing in the presence of misclassification also requires estimation of a form of monotonic function with a known offset. Although, to our knowledge, it has not yet been addressed in the literature, it may be possible that the PAV algorithm could be extended to incorporate a known offset.

As will be seen in Section 3.4, for the regression setting in the presence of covariates, the only link function for which this form of group testing neatly separates to an additive form is the complementary log-log link function.

3.3.4 Grouped Sample with Random Monitoring Times

An alternative group testing scenario which may realistically occur in practice is where the grouped samples are not combined according to a common monitoring time. Instead, samples are randomly combined and then tested as a group. The number of individuals observed at each monitoring time within a group is assumed known, but notpre-defined. For simplicity, assume there exists a single group size k and l unique monitoring times.

Under this group testing scenario, the likelihood of Equation (3.1) can still be written as

$$\prod_{j=1}^{m}[P(\Delta_j = 1|c_j)]^{\delta_j}[P(\Delta_j = 0|c_j)]^{1-\delta_j}.$$

However, the probabilities $P(\Delta_j = 1|c_j)$ and $P(\Delta_j = 0|c_j)$ are now defined as

$$\begin{aligned} P(\Delta_j = 1|c_j) &= P(\Delta_j = 1|y_j = 1, c_j)P(y_j = 1|c_j) + P(\Delta_j = 1|y_j = 0, c_j)P(y_j = 0|c_j) \\ &= \alpha - (\alpha + \beta - 1)(1 - F(c_1))(1 - F(c_2)).....(1 - F(c_k)) \\ &= \alpha - (\alpha + \beta - 1)\prod_{i=1}^{k}(1 - F(c_i)) \end{aligned}$$

and

$$\begin{aligned} P(\Delta_j = 0|c_j) &= P(\Delta_j = 0|y_j = 0, c_j)P(y_j = 0|c_j) + P(\Delta_j = 0|y_j = 1, c_j)P(y_j = 1|c_j) \\ &= 1 - \alpha + (\alpha + \beta - 1)(1 - F(c_1))(1 - F(c_2)).....(1 - F(c_k)) \\ &= 1 - \alpha + (\alpha + \beta - 1)\prod_{i=1}^{k}(1 - F(c_i)) \end{aligned}$$

Assuming $\gamma = \alpha + \beta - 1$, the (conditional) likelihood function allowing for constant misclassification in the response variable with this type of group testing can be written as

$$\prod_{j=1}^{m} [\alpha - \gamma \prod_{i=1}^{k} (1 - F(c_i))]^{\delta_j} [1 - \alpha + \gamma \prod_{i=1}^{k} (1 - F(c_i))]^{1-\delta_j},$$

with the corresponding log-likelihood given by

$$\sum_{j=1}^{m} \delta_j \log(\alpha - \gamma \prod_{i=1}^{k} (1 - F(c_i))) + \sum_{j=1}^{m} (1 - \delta_j) \log(1 - \alpha + \gamma \prod_{i=1}^{k} (1 - F(c_i))).$$

Following the technique of the previous sections, define the distribution function G as

$$G(c_1\, c_2 \,...c_l) = \alpha - \gamma \prod_{i=1}^{k} (1 - F(c_i)),$$

where $(c_1\, c_2 \,...\, c_l)$ are the unique monitoring times found within a specific group, which are monotonic in each argument. A straightforward algorithm for estimation of G is potentially unknown, the PAV algorithm is not appropriate here in general, due to the grouping of different monitoring times.

Recall the time varying misclassification scenario of Section 2.2.3 and the identifiability issue discussed within. A similar identifiability issue can be found here where, even if a reasonable estimator of G is determined, it is not generally possible to solve F directly in terms of G. This issue is most easily seen when there is only a single group, made up of different monitoring times. Suppose this group is determined as positive, impling $\widehat{G} = 1$. Modifying this estimate to allow for the constraints of Equation (3.4) shifts this estimate to $\widehat{G} = \alpha$. Consequently, with $\widehat{G} = \alpha$, this implies $\alpha = \alpha - \gamma \prod_{i=1}^{k} (1 - F(c_i))$. The only solution to this equation is $F(c_i) = 1$ for all monitoring times. This implies all individuals are positive, regardless of the monitoring time. This identifiability issue differs from before in the sense that estimates of F can be identified, however these estimates are not the maximum likelihood estimates. Determining a more appropriate algorithm is therefore of interest.

3.4 Regression

We now extend the group testing scenario (with known misclassification rates) to briefly consider the regression setting under this data structure. The association between standard current status data and generalized linear models was introduced in Section 1.3, along with appropriate references of parametric, semi-parametric and nonparametric regression in the the context of current status data. This regression setting was also discussed in Section 2.2.5

where the correspondence between standard regression models for the underlying failure time and generalized linear models for the observed current status outcome is clearly outlined, both in the presence and absence of misclassification, for simple current status data.

To address regression with group testing and misclassification, first consider the case where there are multiple monitoring times but a single common group size k (Section 3.3.2). In this group testing scenario, using a modified link function, the estimates of the parameters in the regression model of the observed response Δ can be interpreted in terms of the parameters for the unobserved failure time T. To illustrate this result, we follow the approach used in Section 2.2.5 where the regression setting for standard current status data with response misclassification is considered. Again, assume a proportional odds regression model, where a logit link function would be appropriate in the absence of misclassification. Under this setup the distribution function F and the logit link function can be defined as

$$F(c|X) = \frac{1}{1+e^{-a(c)-b(x)}}, \qquad \text{logit link: } \log\left(\frac{p(x|c)}{1-p(x|c)}\right),$$

where $p(x|c) = E[Y|C,X]$. With multiple monitoring times and a single group size k, we define the distribution function $G(c_j) = \alpha - \gamma(1-F(c_j))^k$. Assuming proportional odds regression, G can be described as

$$G = \alpha - \gamma\left(1 - \frac{1}{1+e^{-a(c)-b(x)}}\right)^k = \alpha - \gamma\left(\frac{e^{-a(c)-b(x)}}{1+e^{-a(c)-b(x)}}\right)^k.$$

We now show how misclassification can be incorporated through a simple modification of the link function. The modified link function is now defined as

$$\text{modified link: } \log\left(\frac{1-(\frac{\alpha-p(x|c)}{\gamma})^{1/k}}{(\frac{\alpha-p(x|c)}{\gamma})^{1/k}}\right)$$

This implies

$$g\left\{P(\Delta=1|X,C)\right\} = \log\left(\frac{1-\frac{e^{-a(c)-b(x)}}{1+e^{-a(c)-b(x)}}}{\frac{e^{-a(c)-b(x)}}{1+e^{-a(c)-b(x)}}}\right) = \log\left(\frac{1}{e^{-a(c)-b(x)}}\right)$$
$$= \log(e^{a(c)+b(x)}) = a(c)+b(x)$$

Under this group testing scenario, appropriate modification of other standard link functions will also produce similar results. However, these results do not follow when we extend this analysis to the more complex group testing scenario of Section 3.3.3, where the group sizes vary depending on the monitoring times. It is easily seen from the above calculations that if the groups sizes depend on the monitoring times, the above modified link function cannot be used to describe the required terms additively. Therefore, under the assumed proportional odds regression model, there is no appropriate link function which separates C, X and k_j additively when the group sizes vary with the monitoring times. The only regression setting for which this can be achieved is the proportional hazards regression model, with a complementary log-log link function.

3.5 Discussion

Throughout this chapter we have considered estimation of the distribution function of time to a specific event through the use of group testing, in the presence of misclassification. We develop on the techniques used in Chapter 2 where misclassification of standard current status data is considered. However, alternative methods to allow for misclassification in the current status response variable could also have been considered and modified appropriately. See for example the simulation extrapolation (SIMEX) method, a basic introduction to which is given by Carroll et al. (1996) in the context of simple linear regression. This method has also been applied to binary outcome data with a generalized linear model (Küchenhoff et al., 2006), which could be modified and compared with the results of Section 3.4. This regression setting could also be extended to consider semi-parametric or nonparametric survival models by extending the work of Shiboski (1998a) or Sal y Rosas and Hughes (2011).

Although we have focused primarily on estimating the distribution function F, interest could alternatively be on estimation of some functional of F. In Section 3.3.2 it was seen that in determining the optimal choice of group size k, with multiple monitoring times, examination of a functional may be more informative. Jewell et al. (2006) consider the choice of monitoring mechanism for optimal nonparametric functional estimation for binary data, where it is shown how this optimal choice depends on the functional of interest. The methods of Jewell et al. (2006) could be extended to determine the optimal group size for this form of group testing.

Throughout this chapter we have assumed the misclassification rates are known and do not depend on the group size. As mentioned in Section 2.3, these classification rates may need to be estimated from a validation sample. A similar approach should also be taken to validate the assumption that increasing the group size does not increase the potential for misclassification. Although an application to HIV screening is only appropriate for group testing when estimating the disease prevalence, due to the underlying age-period effect, Kline et al. (1989) show that grouping samples up to a size of $k = 15$ does not alter the levels of sensitivity and specificity from those assumed in the single sample setting. Similar analyses should be performed prior to incorporating specific levels of misclassification for all potential group sizes.

There are many areas in which the ideas of this chapter can be extended to further analyze the use of current status data and group testing in the presence of misclassification as a means to reduce the error rates associated with certain testing mechanisms.

Chapter 4

Current Status Observation of Simple Counting Processes

4.1 Introduction

Current status data is a type of survival data where instead of observing exact failure times we now only observe information on the survival status of individuals at a specific point in time. Chapter 1 gives a review of methods, techniques, and limitations seen with current status data, where the current status framework is introduced through motivating examples (Section 1.1.1), followed by a formal introduction defining the necessary notation (Section 1.1.2). Various sampling schemes that deviate from the simple current status data structure are presented in Section 1.4 and other areas of interest with current status data, including regression (Section 1.3), cure models (Section 1.6) and mortality differentials (Section 1.10), are also addressed within Chapter 1. Subsequent chapters then consider current status data where the current status response variable is subject to misclassification, both when samples are tested individually (Chapter 2) and combined and tested as a group (Chapter 3). As was discussed in Section 1.5, current status data can also be viewed as a counting process. This chapter extends the simple counting process described by Equation (1.2) and considers current status observation of a three-state counting process, which is motivated by an application to simultaneous accurate and diluted HIV test data. For more information on the HIV virus, including prevalence, diagnosis and prevention, see Appendix C.

Recall from Section 1.5.1 that multistate models (Anderson and Keiding, 2002) are a type of multivariate survival data, used to explain how individuals move through a succession of stages corresponding to distinct states. In medicine, the states can describe conditions such as the health status of patients, for example, healthy, diseased, diseased with complication, and deceased could define four such states. Multistate models can also be used when examining the lifetime accumulation of events, such as the lifetime number of sexual partners, marriages, or employment history. Multistate models can also be used in analyzing wear patterns in industrial applications. Equation (1.2) describes the simplest multistate model consisting of at most one event. This simple example could describe the mortality model for survival data, made up of only two states, an initial state (alive) and a terminal state (dead). Multistate models are of course designed to handle more complex situations which allow for more than two stages. A special case of multistate model is the illness-death model which has been studied for right censored data (Anderson et al., 1993, Datta et al., 2000). This model has also been referred to as the disability model. This illness-death model involves three states, the alive and healthy state (State 0), the alive and diseased state (State 1), and the deceased state (State 2). It is often assumed that recovery is not possible so that transition from State 0 to State 1 is irreversible.

A variation of this model is another three state model where the states are defined by the onset and diagnosis of a particular disease. In this chapter we consider this model where the disease is assumed to be irreversible. Irreversibility is necessary as otherwise a negative disease status at the monitoring time may be indicative of either no prior occurrence or occurrence followed by recovery. This approach would therefore not be appropriate for the application to HPV infection used in Chapter 2, due to the high rate of recovery with this specific virus. With HPV infection, it would be possible to move from State 0 to State 1 once infected and then back to State 0 after recovery. The model we consider in this chapter is a progressive three-state model which implies the only possible transitions are from State 0 $\rightarrow$ State 1 $\rightarrow$ State 2. We also assume that once an individual is a member of State 1, they will eventually transition to State 2. This assumption is reasonable for many applications, including our application to HIV testing where a positive test result on a sensitive test implies a positive result on a less sensitive test will eventually occur. However, in certain applications this assumption may not hold, for example, when assessing the lifetime accumulation of events, where having a first job does not necessarily imply later having a second job.

The complexity of multistate models depends on the number of possible states and by the manner in which individuals are allowed to transition between these states. Figure 4.1 shows different multistate models, the complexity of which increases from Model (a) to Model (d). In this chapter we consider the straightforward progressive three-state model illustrated by Model (a) in Figure 4.1. The ideas we present here can also be extended to more complex models, such as those illustrated in Model (b) and Model (c) of Figure 4.1, both of which also represent progressive models where the transitions move through a succession of stages. The methods we consider in this chapter cannot be applied to a model such as an illness-death model with recovery, as displayed in Model (d) of Figure 4.1, due to its cyclic nature where recovery is represented by the double sided arrow between State 0 and State 1.

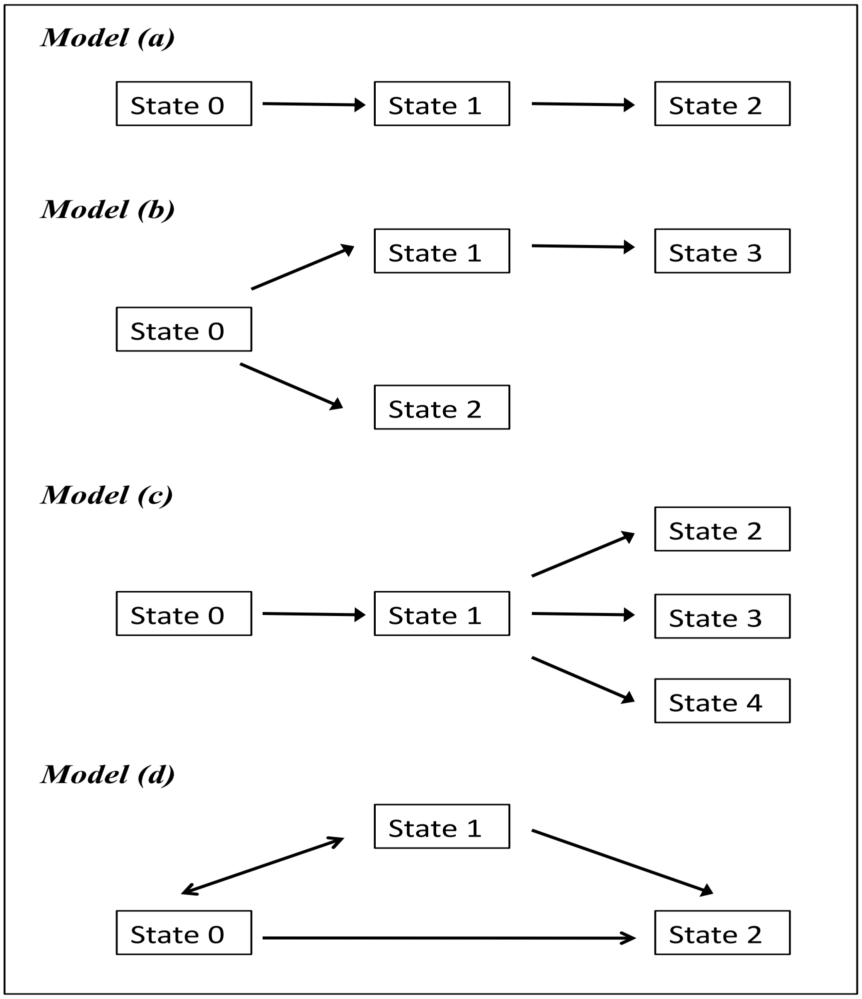

Figure 4.1: Illustration of four multistate models, the complexity of which increases from Model (a) to Model (d).

As is expected with survival data, we may not see complete information for each individual due to censoring. Multistate data can be observed for right censored data (Meira-Machado et al., 2008), interval censored data (Sutradhar et al., 2010) or even current status data (Datta and Sundaram, 2006, Lan and Datta, 2008). In multistate current status data one only observes whether or not each of the individual survival times defining the different stages exceed the common observed monitoring time. In this chapter we examine multistate current status data in terms of counting processes, introduced for the simple case in Section 1.5. For ease of explanation and understanding, throughout this chapter, we focus on the special multistate case consisting of only two jumps, or events of interest, but the ideas can easily be extended to allow for more jumps. Practical examples of this kind of data structure, with only two possible jumps, arise in a number of applications, including carcinogenicity experiments (van der Laan et al., 1997b), examining the onset and diagnosis of uterine fibroids (Dunson and Baird, 2001, Young et al., 2008), and the detection of HIV infection through accurate and diluted assays (Balasubramanian and Lagakos, 2010). Additional applications can be found in Jewell and van der Laan (1995). The motivation behind this work relates to the application to HIV testing, more information on which is given in Section 4.4 with a detailed overview of HIV given in Appendix C.

The main focus of this chapter is estimation of the distribution function of time to the first event (or in terms of counting processes, the time until the first jump). Interest could alternatively be on estimation of time until the second event, where similar techniques would apply. With the application to HIV testing considered in Section 4.4, knowledge of the time until to first event is of greater practical benefit. More specifically, given multistate current status data, we examine whether current status information on one event can be used to improve the estimate of the distribution function of time to the other event. In particular, we focus on estimation of the distribution function of time to the first event and whether current status information on a subsequent event can improve this estimate. Much attention has been given to a similar progressive three-state model consisting of a right censored observation of the final event with a current status observation of the intermediate event. For this related data structure it has been shown that in the fully nonparametric setting, one cannot improve the naïve Kaplan-Meier estimator when estimating smooth functionals of the distribution of times to the first event in most situations (van der Laan and Jewell, 2003). Nonparametric estimation in the single-sample setting, along with simple alternatives are discussed by van der Laan et al. (1997b). Furthermore, under this data structure, regression models for time to the first event are considered by Dunson and Baird (2001) and Young et al. (2008), under proportional hazards.

Returning to the multistate current status scenario of interest, we now extend the results of these related models to consider the pure current status case where we only observe current status information on both the final and intermediate events. In van der Laan and Jewell (2003) it is shown that in the fully nonparametric setting, one cannot improve the naïve current status estimators when estimating smooth functionals of the distribution of time to the first, or second, event. A naïve current status estimator used to examine the time until a particular event can be defined as an estimator that uses current status information on the event of interest only, and does not use any information on subsequent events. For our particular case, since we are interested in estimating the time until the first event, our naïve current status estimator would therefore only use information on the current status of the first event and would ignore any additional information provided by the second event. For the naïve estimator, nonparametric estimation in the single-sample setting can then be obtained through the well-known pool-adjacent violators (PAV) algorithm, described in detail in Section 1.2. Analysis of the time to a specific event is often also of interest in a semi-parametric setting. For a three-state counting process, the semi-parametric setting may be obtained by making parametric assumptions about the time to the first event, without making any assumptions about the time to the second event, and vice-versa. Alternatively, parametric assumptions may be made about the length of time between the two events, with the distributions of time to the first and second events left unspecified. For convenience, throughout this chapter, we refer to this time between the two events as the waiting time.

As stated by van der Laan and Jewell (2003), for the related three-state model consisting of both right censored and current status observations, when no parametric assumptions are made, the naïve estimator of time to the first event cannot be improved by including additional information on other events. For completeness, in Section 4.2.1 we show that this result also follows through to our case with multistate current status data, as expected. We then examine whether one can improve these naïve estimators when parametric assumptions about the time between the two events are made. Using assumptions about the waiting time between events, we estimate the nonparametric maximum likelihood estimate (NPMLE) of the distribution of time to the first jump, or time at which the first event occurs. Later, in the illustration of Section 4.4, we also consider estimation of the cumulative hazard of the distribution of time to the first event over a short interval prior to the monitoring time. This corresponds to estimating this cumulative hazard in the recent past. Finally, for situations where the waiting time between the two events is modifiable by design, as is possible in an application such as that using accurate and diluted HIV assays, we also address the issue of determining the optimal length of the waiting time for estimation of this cumulative hazard. In Section 4.4 we illustrate the ideas and methods of this chapter through application to simultaneous accurate and diluted HIV test data.

4.2 Notation and Likelihood

We now extend the notation of Section 2.2 for simple current status data to describe our multistate models. For simplicity, we focus on a special case of counting process where there are at most two jumps and we obtain current status observation of this counting process. Again, the methods used can easily be extended to allow for more states, or to incorporate more complex multistate models such as those shown in Models (b) and (c) of Figure 4.1.

Let T_1 and T_2 be the survival time random variables of interest for the time to the first and second events, respectively. Hence, $T_1 \leq T_2$. Note that this setup differs from the bivariate current status data structure of Section 1.4.5 due to this ordering of the events. This condition implies the first event will always occur before (or at least at the same time as) the second event, a concept that will have importance later in this chapter. The corresponding distribution functions (survival functions) for the first and second events are F_1 (S_1) and F_2 (S_2), respectively. Keeping consistent with the previous chapters, let the monitoring time be denoted by the random variable C. For convenience we describe the random monitoring time scenario, where current status observation now refers to a sampling scheme where n

i.i.d. observations are collected on the random variable (Δ_1, Δ_2, C), where $\Delta_1 = I(T_1 \leq C)$ and $\Delta_2 = I(T_2 \leq C)$. Assume the monitoring time C is independent of T_1 and T_2 and focus on the conditional likelihood, given C. Under this data structure, there are three possible types of observations

(1) $\Delta_1 = 0$ and $\Delta_2 = 0$ when $T_1 > C$ and $T_2 > C$

(2) $\Delta_1 = 1$ and $\Delta_2 = 0$ when $T_1 \leq C$ and $T_2 > C$

(3) $\Delta_1 = 1$ and $\Delta_2 = 1$ when $T_1 \leq C$ and $T_2 \leq C$.

Figure 4.2 gives a schematic of the way in which individuals can progress through the different states defined by the occurrence of events at T_1 and T_2. At the monitoring time C, an individual is defined as a member of State 0 if $\Delta_1 = 0$ and $\Delta_2 = 0$, a member of State 1 if $\Delta_1 = 1$ and $\Delta_2 = 0$, and a member of State 2 if $\Delta_1 = 1$ and $\Delta_2 = 1$. Let n_{00}, n_{10}, n_{11} be the total number of individuals observed in each of these three states.

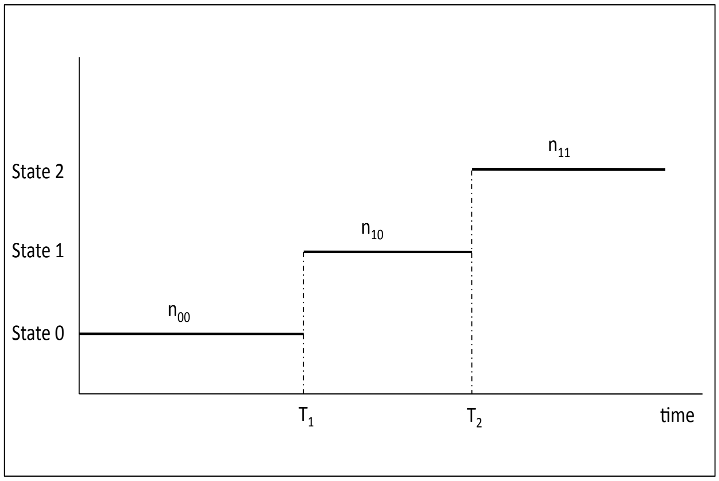

Figure 4.2: Schematic of the progression of disease through the states defined by the occurrence of events at T_1 and T_2. The number of individuals in States 0, 1, 2 are given by n_{00}, n_{10}, n_{11}, respectively.

This data structure can be viewed as current status observation of a two-jump (or three-state) counting process. A general counting process with k jumps is defined by

$$N(t) = \sum_{j=1}^{k} I(T_j \leq t), \quad \text{with } T_1 < ... < T_k,$$

where T_j is the time variable at which the jth event occurs, and k is the maximum number of possible events. When an event occurs the value of N increases by one unit. Therefore, at the time of the j^{th} event, N will jump from $j-1$ to j. Throughout this chapter we assume the maximum number of possible jumps is known and fixed at $k=2$. The assumption that C is independent of T_1 and T_2 translates to assuming C is independent of N. Under this data structure the three possible observations for N can be summarized as

$$N(C) = \begin{cases} 0 & \text{if} \quad C < T_1 \\ 1 & \text{if} \quad T_1 \leq C < T_2 \\ 2 & \text{if} \quad T_2 \leq C \end{cases}.$$

Note this is a straightforward extension of the counting process described in Section 1.5.

4.2.1 Likelihood

To define the likelihood for this data structure, let C_i be the i^{th} order statistic of $C_1, C_2, ..., C_n$. Let δ_{00i} be an indicator variable counting if the i^{th} observation has $\Delta_1 = 0$, $\Delta_2 = 0$. Similarly, δ_{10i} and δ_{11i} define the indicators counting if the i^{th} observation has $\Delta_1 = 1$, $\Delta_2 = 0$ and $\Delta_1 = 1$, $\Delta_2 = 1$, respectively. The corresponding (conditional) likelihood function can then be given by

$$\prod_{i=1}^{n} (1 - F_1(C_i))^{\delta_{00i}} (F_1(C_i) - F_2(C_i))^{\delta_{10i}} F_2(C_i)^{\delta_{11i}}, \tag{4.1}$$

where F_1 and F_2 are the distribution functions of T_1 and T_2, respectively. Since this likelihood can be fully expressed in terms of the marginal distributions of T_1 and T_2, the only identifiable parameters are F_1 and F_2. The constraint that the second event cannot occur before the first event implies that $P(T_1 \leq T_2) = 1$. This constraint can equivalently be written in terms of the marginal distributions, namely that $F_1(t) \geq F_2(t)$. This condition holds because for any pair of marginal distributions (F_1, F_2) with $F_1 \geq F_2$, there exists a bivariate distribution with F_1 and F_2 as the marginals with the corresponding $P(T_1 \leq T_2) = 1$. We subsequently perform our analyses based solely on the marginal distributions, F_1 and F_2. The ideas of this section can be extended to allow for covariate effects on F_1, F_2, or both. Under the proportional hazards assumption, regression models have been considered by Dunson and Baird (2001) and Young et al. (2008) for the related multistate model.

Using the ideas of Jewell et al. (2003), originally applied to competing risks current status data (see Section 1.7), we re-parameterize the likelihood in terms of the ratio of the survival functions. The likelihood given in Equation (4.1) can then be written as

$$\prod_{i=1}^{n} R(C_i)^{\delta_{00i}} (1 - R(C_i))^{\delta_{10i}} S_2(C_i)^{\delta_{00i}+\delta_{10i}} (1 - S_2(C_i))^{\delta_{11i}}, \tag{4.2}$$

where the ratio of the survival functions is defined by

$$R(C_i) = \frac{S_1(C_i)}{S_2(C_i)} = P(T_1 > C_i \mid T_2 > C_i).$$

The corresponding log-likelihood is

$$\sum_{i=1}^{n} \delta_{00i} \log(R(C_i)) + \sum_{i=1}^{n} \delta_{10i} \log(1 - R(C_i))$$
$$+ \sum_{i=1}^{n} (\delta_{00i} + \delta_{10i}) \log(S_2(C_i)) + \sum_{i=1}^{n} \delta_{11i} \log(1 - S_2(C_i)). \qquad (4.3)$$

Note that $S_1 = S_2 R$ must be a survival function and that $P(T_1 \leq T_2) = 1$. Therefore, R and S_2 are functionally dependent. Without this consideration, maximization of Equation (4.3) may result in an estimate for S_1 which is not a survival function.

Suppose that all individuals are examined at the same monitoring time C. In some applications it will also be necessary to assume that all individuals are the same age at this time of testing, due to the impact of an underlying age-period effect. This assumption is made in the illustration of Section 4.4. For this case of a single monitoring time for all individuals, the re-parameterized likelihood of Equation (4.2) can then be written as

$$R(C)^{n_{00}} (1 - R(C))^{n_{10}} S_2(C)^{n_{00} + n_{10}} (1 - S_2(C))^{n_{11}}, \qquad (4.4)$$

where n_{00}, n_{10}, n_{11} are the total number of individuals observed in State 0, State 1, State 2, respectively. Under this single monitoring time scenario, the NPMLE of either S_1 or S_2 can be obtained through the use of naïve current status estimators, maximizing this likelihood (Equation (4.4)). To obtain each naïve estimate, we focus solely on the jump of interest and use standard maximization techniques. The straightforward PAV algorithm can be used directly to obtain each of these estimates. Applying the PAV algorithm to $(1 - \Delta_2)$ results in the NPMLE of $S_2(C)$. Similarly, applying this algorithm separately to $(1 - \Delta_1 | \Delta_2 = 0)$ results in the NPMLE of $R(C)$. Consequently, $S_1(C)$ is easily obtained through the definition, $S_1(C) = R(C)S_2(C)$. In fact, for the single monitoring time setting, the nonparametric estimates of $S_1(C)$, $S_2(C)$ and $R(C)$ can be determined by direct maximization of the likelihood in Equation (4.4), the results of which can be written in terms of n_{00}, n_{10}, and n_{11} without the need for an algorithm. These maximum likelihood values can be found through the following simple definitions

$$\hat{S}_1(C) = \left(\frac{n_{00}}{n}\right), \quad \hat{S}_2(C) = \left(1 - \frac{n_{11}}{n}\right), \quad \hat{R}(C) = \left(\frac{n_{00}}{n_{10} + n_{00}}\right).$$

These results imply that for a single monitoring time, it is not possible to improve the naïve current status estimates, when no assumptions are made about the distributions of T_1, T_2 or the waiting time between the two events. This is evident from the above definition where the NPMLE of $S_1(C)$ is based solely on current status information on the first event. Note that n_{00} is the number of individuals for whom neither event has occurred. Recall the ordering of events, if the first event has not occurred by time C, the second event has also not yet occurred. Therefore, all the required information to estimate the NPMLE of $S_1(C)$ can be obtained through current status information on the first event.

We now consider the case where multiple monitoring times exist. This is represented by the likelihood in Equation (4.2) where $R(C_i)$ is the ratio of survival functions for the i^{th} individual. This likelihood is slightly more complicated than that of Equation (4.4). In this more complicated case the NPMLE of S_1 cannot be obtained directly using the PAV algorithm. However, note that the definition of $R(C_i)$ can be written as

$$R(C_i) = S_1(C_i)/S_2(C_i) = E(1 - \Delta_1 | C = C_i, T_2 > C_i).$$

The definition of $S_1(C_i)$ can then be written as

$$\begin{aligned} S_1(C_i) = S_2(C_i)R(C_i) &= S_2(C_i)E(1 - \Delta_1 | C = C_i, T_2 > C_i) \\ &= E\{S_2(C_i)(1 - \Delta_1) | C = C_i, T_2 > C_i\}. \end{aligned}$$

Although the naïve estimator cannot be obtained using the PAV algorithm, a 'sophisticated naïve' approach exists, similar to the ad hoc approach of Jewell et al. (2003), outlined in Section 1.7. As stated by van der Laan et al. (1997b), estimating S_1 can be viewed as estimating a monotonic regression of $S_2(C_i)(1-\Delta_1)$ on the observed monitoring times. This approach suggests replacing $S_2(C_i)$ by the pool-adjacent violators estimate of $S_2(C_i)$ and then minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\{S_2(C_i)(1 - \Delta_{1i}) - S_1(C_i)\}^2 I(C_i \leq T_{2i})v_i$$

over the vector $\{S_1(C_i) : i = 1, ..., n\}$, subject to the constraint that S_1 is monotone. Here v_i represents the weights for individual i, with this weight defined by

$$v_i = \frac{1}{S_2^2(C_i)R(C_i)(1 - R(C_i))}.$$

This maximization can be achieved through a weighted pool-adjacent violators algorithm, which is described in Dinse and Lagakos (1982). These results show that in the fully nonparametric setting, we cannot improve the naïve current status estimators when estimating F_1 (or F_2), regardless of whether there are single or multiple monitoring times. This result is as expected and agrees with the results found in van der Laan and Jewell (2003) relating to the multistate model consisting of a right censored observation of the final event and a current status observation of the intermediate event.

4.3 Waiting times, W

Let W be the waiting time random variable between the first and second events. This waiting time is simply defined as the length of time between the two events, $W = T_2 - T_1$. Figure 4.3 gives a schematic identifying the variables T_1, T_2 and W. This schematic can be viewed as a progression through time, beginning at the timing of the first event. The solid arrow (top) represents a positive outcome for the first event ($\Delta_1 = 1$), which begins at time T_1. The dotted arrow (bottom) begins with the occurrence of the second event at time T_2, after which $\Delta_2 = 1$. The time during which these arrows do not overlap is considered the waiting time, W. As the main focus of this chapter is estimation of the time until the first event, hereafter we refer to the timing of the second event, T_2, as $T_1 + W$.

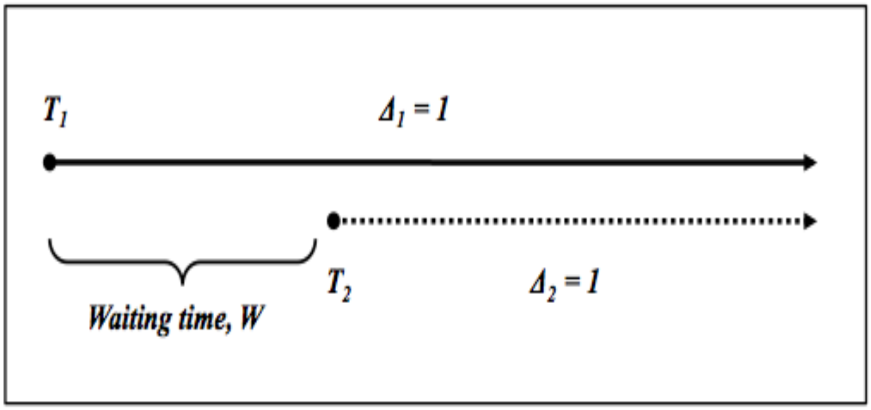

Figure 4.3: Schematic illustrating the definition of the waiting time variable, W in relation to the occurrence of events at times T_1 and T_2.

From Section 4.2.1 we see that since the likelihood only depends on the marginal distributions of T_1 and T_2, one cannot make inferences on W without further assumptions, for example, that W and T_1 are independent. In van der Laan and Jewell (2003) it is seen that at many continuous data generating distributions the PAV algorithm estimates for the marginal distributions of the unobservable time variables yield asymptotically efficient estimators of $\sqrt{n}$-estimable parameters. We now make some assumptions about the waiting time random variable, W, in order to potentially gain from the additional information on the second event. Specifically we examine the possibility of improving the naïve estimators when parametric assumptions about the waiting time between the two events are made.

Again, we assume the disease of interest is irreversible. We also make the assumption that W has support in the interval $[0, W^*]$ and that $W^* \leq C$, which is true in most practical settings. The assumption of a finite support implies that if the first event occurs, the second event is guaranteed to eventually occur. As mentioned in Section 4.1, this assumption may not hold in examples such as examining the lifetime accumulation of events, where the occurrence of one event does not necessarily imply the eventual occurrence of a subsequent event. However, this assumption is valid in many medical and public health applications, particularly those related to the onset and diagnosis of disease.

In the remainder of this chapter we explore how various assumptions made about this waiting time W impact our estimates of F_1, both mathematically and through applications to simulated data. We address how the likelihood and maximization of this likelihood change depending on whether these waiting times are known or unknown and fixed or varying. In Section 4.3.1 we assume there is a single fixed waiting time for all individuals. We then relax this constraint in Section 4.3.2 to consider the case where the waiting times may differ between individuals, and identify the algorithmic issues faced in such a scenario. Assuming a single monitoring time, we also consider estimation of the cumulative hazard of F_1 over a short period prior to C, that is, the cumulative hazard of F_1 in the recent past. Finally, in Section 4.4 we examine the impact of the assumption that F_1 is increasing linearly in the interval $[C - W^*, C]$ through application to simultaneous accurate and diluted HIV test data, an example which is examined by Balasubramanian and Lagakos (2010).

The introduction of the waiting time random variable W gives the following likelihood for a single monitoring time C

$$\prod_{i=1}^{n}(1 - F_1(C))^{\delta_{00i}}(F_1(C) - F_1(C - w_i))^{\delta_{10i}}F_1(C - w_i)^{\delta_{11i}}, \tag{4.5}$$

where δ_{00i}, δ_{10i}, δ_{11i} are indicators counting if the ith observation has $\{T_1 > C, T_2 > C\}$, $\{T_1 < C, T_2 > C\}$, $\{T_1 < C, T_2 < C\}$, respectively, and w_i is the observed value of W for individual i. As before, this likelihood can also be re-parameterized in terms of the ratio of the distribution functions, or equivalently, the ratio of the survival functions. The likelihood of Equation (4.5) can therefore be written as

$$\prod_{i=1}^{n} R_i(C)^{\delta_{11i}}(1 - R_i(C))^{\delta_{10i}}F_1(C)^{\delta_{10i}+\delta_{11i}}(1 - F_1(C))^{\delta_{00i}},$$

where

$$R_i(C) = \frac{F_1(C - w_i)}{F_1(C)} \quad \text{and} \quad 0 \leq R_i(C) \leq 1.$$

Again, estimation of the NPMLE of F_1 is of interest, where ideally the waiting times are random and can potentially differ for each individual. To address maximization of this likelihood (Equation (4.5)) we first consider the simplest waiting time scenario where the waiting time does not vary between individuals. We then consider the multiple waiting time scenario and specifically address the special case where all individuals within a state have the same waiting time. We use this scenario to highlight the problems faced when considering the purely random waiting times. The following sections consider these scenarios and also address the case where the waiting times are unobserved but their distribution is known.

4.3.1 Single Waiting Time

First, assume there is a true single waiting time, W, which is known and fixed regardless of the number of events that have occurred for each individual. Under this scenario, the likelihood of Equation (4.5) can be simplified to

$$(1 - F_1(C))^{n_{00}}(F_1(C) - F_1(C - W))^{n_{10}}F_1(C - W)^{n_{11}}, \tag{4.6}$$

again where n_{00}, n_{10}, n_{11} are the number of observations in States 0, 1, 2, respectively. This likelihood is equivalent to an interval censored problem where each individual belongs to exactly one of the intervals, $(C, \infty]$, $(C - W, C]$, $(0, C - W]$, corresponding to State 0, State 1, State 2, as previously described. Maximization of this likelihood can then be obtained through the use of standard interval censored methods. For a review of interval censored data and the appropriate methods of estimation, see Zhang and Sun (2010). For a single monitoring time C and a single waiting time W, estimates of $F_1(C-W)$ and $F_1(C)$ can easily be obtained through an Expectation Maximization Iterative Convex Minorant (EM-ICM) algorithm (Sun, 2006). In this instance, the algorithm converges in a single iteration. Due to the simplicity of the likelihood in Equation (4.6), estimates of $F_1(C - W)$ and $F_1(C)$ can also be obtained directly through the following definitions

$$\hat{F}_1(C - W) = \frac{n_{11}}{n} \quad \text{and} \quad \hat{F}_1(C) = 1 - \frac{n_{00}}{n}.$$

If multiple monitoring times exist, yet there is still only a single known waiting time W for all individuals, we could use the ‘sophisticated naïve’ method outlined in Section 4.2.1, based on isotonic regression against C to estimate R. In this situation, the ratio of the distribution functions for individual i would be defined by

$$R(c_i) = \frac{F_1(c_i - W)}{F_1(c_i)}.$$

Therefore, for a single fixed waiting time, estimates of both $F_1(c_i - W)$ and $F_1(c_i)$ are easily obtained, regardless of the number of unique monitoring times.

4.3.2 Random Waiting Times

We now move from this simplistic scenario and consider the case where the waiting times may vary across individuals. Ideally we would like to consider the random waiting time scenario where there could potentially be a different waiting time for each individual, a scenario which may realistically arise in practice. However, in this setting a straightforward algorithm for estimation of the NPMLE of F_1 is not obvious to us without making further assumptions about the distribution function F_1. In this section we highlight the complications faced in obtaining such an algorithm. In Section 4.4 we show that these algorithmic issues become tractable when certain assumptions about F_1 are made. We then use these assumptions to consider estimation of the cumulative hazard of time to the first event in the recent past. The results of which are presented for simulated data in Section 4.4.

To illustrate the issues associated with examining the random waiting time scenario, without making additional assumptions on F_1, consider the case where all individuals within a specific state have the same waiting time. These waiting times can then vary from state to state but not from individual to individual within the same state. Let w_0, w_1, w_2 be the waiting times for all members of State 0, State 1, State 2, respectively. This means, for individual i, w_0 occurs when $\delta_{00i} = 1$, w_1 occurs when $\delta_{10i} = 1$ and w_2 occurs when $\delta_{11i} = 1$. An individual can only belong to a single state and therefore only one of the defined waiting times will be applicable to each individual. The likelihood for a single monitoring time with this specific multiple waiting time scenario can be written as

$$(1 - F_1(C))^{n_{00}}(F_1(C) - F_1(C - w_1))^{n_{10}} F_1(C - w_2)^{n_{11}}, \tag{4.7}$$

where n_{00}, n_{10}, n_{11} are the total number of individuals in State 0, State 1, State 2, respectively. This likelihood can also be re-parameterized as

$$R_2(C)^{n_{11}}(1 - R_1(C))^{n_{10}} F_1(C)^{n_{10}+n_{11}}(1 - F_1(C))^{n_{00}},$$

where

$$R_1(C) = \frac{F_1(C - w_1)}{F_1(c)}, \quad R_2(C) = \frac{F_1(C - w_2)}{F_1(c)},$$

$$0 \leq R_1(C) \leq 1, \quad 0 \leq R_2(C) \leq 1.$$

The likelihood in Equation (4.6) is a special case of the likelihood in Equation (4.7), where $w_0 = w_1 = w_2 = W$, thereby simplifying maximization of the likelihood. When the waiting times are not equal across states, it is important to note that the length of the waiting time in State 1 (w_1) relative to the length of the waiting time in State 2 (w_2) will impact maximization of this likelihood in Equation (4.7). If the waiting time in State 1 is shorter than that of State 2 ($w_1 < w_2$) it follows that $R_1(C) \geq R_2(C)$, and estimating $R_1(C)$ and $R_2(C)$ reduces to a simple interval censoring problem. The NPMLE of F_1 can then be obtained using an EM-ICM algorithm, as introduced in Section 4.3. However, if the reverse is true, with $w_1 > w_2$, it follows that $R_1(C) \leq R_2(C)$. Estimation of $R_1(C)$ and $R_2(C)$ then becomes a more complex interval censoring problem. Figure 4.4 highlights the difference between these two interval censoring problems, the complexity of which depends on the value of w_1 relative to w_2. Figure 4.4(a) represents the simpler interval censoring problem to which standard methods can be applied. The more complex scenario with overlapping intervals is given in Figure 4.4(b). Each plot can be viewed as a progression through time, beginning at time 0, where the monitoring time C is assumed to be out to the right hand side. The waiting time for State 0 (w_0) is not shown in either plot as it is the waiting time for an individual for whom neither event has occurred, and knowledge of $F_1(C - w_0)$ is not necessary for maximization of the likelihood in Equation (4.7).

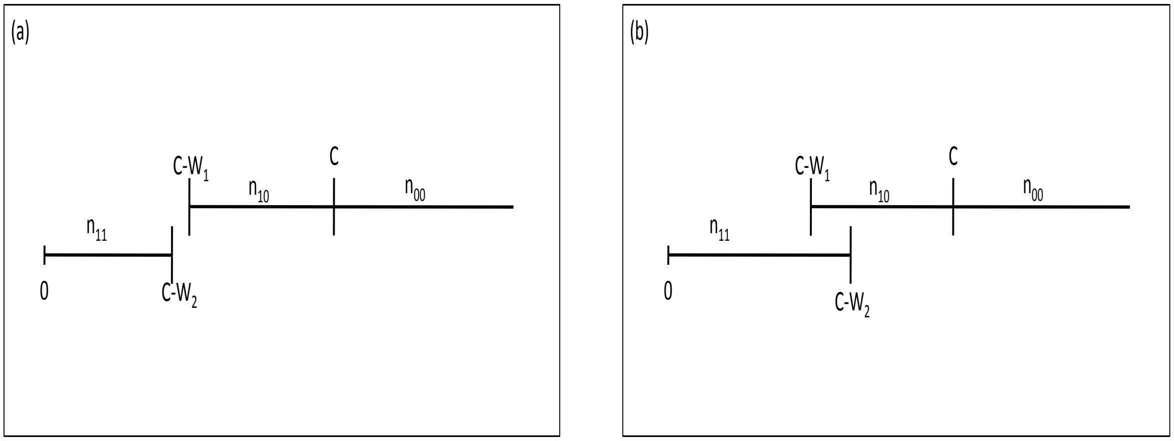

Figure 4.4: (a) Interval censoring problem when $w_1 < w_2$. (b) Interval censoring problem when $w_1 > w_2$. Schematic of interval censoring problems under different scenarios, which depend on the values of the waiting times in State 1 and State 2, w_1 and w_2.

The Expectation-Maximization (EM) algorithm of Dempster et al. (1977) is a well-known iterative algorithm used to obtain the maximum likelihood estimates of parameters in a model that depends on unobserved latent variables. Although this algorithm seems appropriate for our particular setting, a naïve EM algorithm, where the augmented data adds the unobserved W values, is in fact not always appropriate. Using such an algorithm to estimate $F_1(C-w_1)$, $F_1(C - w_2)$ and $F(C)$ we obtained the following results.
If $w_1 < w_2$, the maximum likelihood solution obtained is

$$\hat{F}_1(C - w_1) = \hat{F}_1(C - w_2) = \frac{n_{11}}{n} \quad \text{and} \quad \hat{F}_1(C) = 1 - \frac{n_{00}}{n}.$$

If $w_1 > w_2$, the maximum likelihood solution obtained is

$$\hat{F}_1(C - w_1) = 0 \quad \text{and} \quad \hat{F}_1(C - w_2) = \hat{F}_1(C) = \frac{n_{10} + n_{11}}{n}.$$

The implication of this result is such that even when the waiting time distribution is known, but W itself is not observed, this simplistic EM algorithm is not appropriate. Regardless of the actual distribution of W, the EM algorithm always results in $w_1 > w_2$. The EM algorithm alternates between an E-step (expectation) and an M-step (maximization). In the E-step, the expectation of the log-likelihood function is calculated based on the current estimates of the latent variables, in this case, the unobserved waiting times. The M-step then computes the parameters maximizing the expected log-likelihood obtained through the E-step. These estimates are then used in the next iteration of the E-step. The algorithm continues until a predefined level of convergence is attained. See Dempster et al. (1977) for a more complete understanding of the EM algorithm. Note that individuals with large values of W are more likely to be detected as members of State 1 than of State 2. If a large waiting time exists, the chance of being monitored during this waiting time increases, at which time an individual would be defined as being a member of State 1. If the waiting time is short, an individual is less likely to be monitored during this waiting time, and would instead be observed when both events have already occurred, placing that individual in State 2. For this reason, this naïve EM algorithm is not a suitable algorithm for the random waiting time scenario, even in this specific scenario with a single waiting time for each state. A more appropriate approach would potentially base the augmented data on the score function. However, such an algorithm is not trivial for an unknown F_1 distribution, even when the waiting times are subject to a known distribution. This remains a challenging problem.

Wang and Lagakos (2010) address this issue of length-biased sampling by proposing an augmented design structure whereby individuals in State 1 are followed longitudinally for a more accurate estimate of the time at which the second event occurs, T_2. In many applications this will be a feasible modification, especially if the disease of interest is rare, and the necessary length of follow-up is not too long. However, this augmented design consideration deviates from the current status data structure addressed here, where such additional information could not be obtained. We focus on the pure multistate current status observations but recognize the benefits of the augmented design approach, particularly when only a small proportion of individuals are found to be in State 1, as would be likely for certain infectious disease studies.

Due to the limitations addressed above, when the waiting times are unknown, it is therefore helpful to make an additional assumption to make the likelihood more tractable. One such assumption could be that the distribution function F_1 is increasing linearly over a short interval prior to the monitoring time, say over the interval $[C - W^*, C]$. Equivalently, we could assume the corresponding density $f_1(t) = f$ for t in the interval $[C - W^*, C]$, see Balasubramanian and Lagakos (2010) in their application to HIV incidence estimation. Using this additional assumption, the following section examines simulated data which is motivated by an application to simultaneous accurate and diluted HIV test data. We also retain the assumptions of the previous sections, namely that the disease of interest is irreversible, W has finite support and W is independent of T_1.

4.4 Illustration

Recent advances in the area of HIV incidence estimation motivate and illustrate this work. One such advancement was given by Janssen et al. (1998), who made use of the fact that HIV infection is followed within a few weeks by the development of HIV specific antibodies. This led to the development of a now popular alternative approach for estimating HIV incidence rates in which each subject is simultaneously given both a sensitive diagnostic test (accurate) and a less-sensitive diagnostic test (diluted). Those found with a positive result on the first test and a negative result on the second test are considered 'recently infected'. For a general understanding of HIV infection, its prevalence, biological features and methods of detection and prevention, see Appendix C. To understand the basis of this approach to identifying recent infections, consider the basic schematic given in Figure 4.5 which shows HIV progression over time.

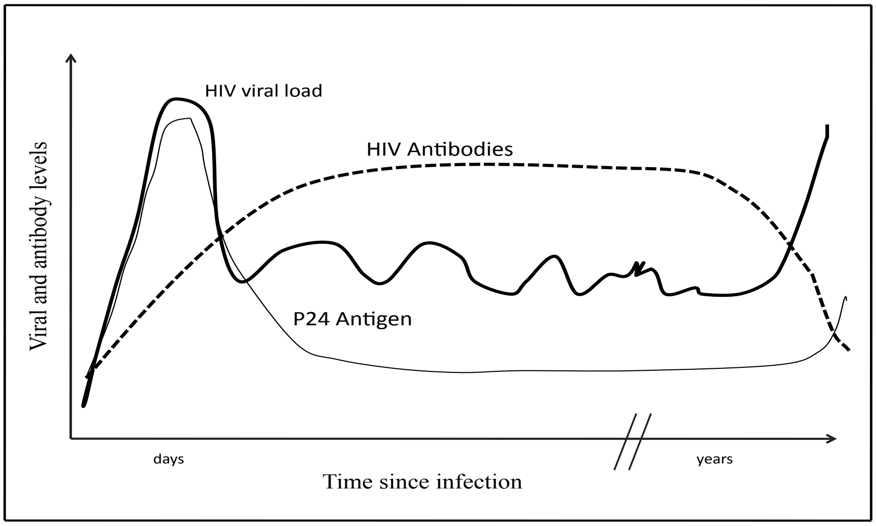

Figure 4.5: Schematic of HIV progression over time.

Figure 4.5 shows the evolution of biomarkers over time where the time scale begins at the time of HIV infection. The simple interpretation of this figure is that the introduction of the P24 antigen causes HIV specific antibodies to develop to neutralize this antigen. The detection of certain biomarkers can be used to indicate a particular disease state. We can see from this diagram that the level of antibodies present for someone who has only been infected for a few days is quite different from the antibody levels of someone who has been infected for several years. Differentiating between individuals with early HIV infection and those infected for longer periods is difficult but important in estimating HIV incidence, and for purpose of clinical care and prevention. If you test for a very high antibody level, only those who have been infected for a long period of time will show a positive result. If you test for a very low antibody level, almost all of the infected individuals will appear positive. It is this feature that Janssen et al. (1998) exploit to identify recently infected individuals. It is important to note that, due to the nature of the development of antibodies, a negative result on an antibody test may not necessarily indicate no infection, it simply indicates there are not yet enough detectable antibodies. We can see from Figure 4.5 that soon after infection, false negatives may be produced. In this chapter, we are interested in estimating the time until a positive result on the sensitive test, and not specifically the time until infection. However, the results of Chapter 2 could be applied if estimation of the time until infection were of interest and this type of false positive needed to be considered.

In a recent paper, Balasubramanian and Lagakos (2010) use the results of Janssen et al. (1998) to estimate HIV incidence for HIV test data obtained through the use of sensitive and less-sensitive diagnostic tests. The standard ELISA antibody assay was used for the sensitive test and a detuned version of the ELISA was used for the less-sensitive test. More information on the ELISA (Enzyme-Linked ImmunoSorbent Assay) is given in Appendix C. This type of data is consistent with the data structure outlined in Section 4.2 and can be viewed from a current status data perspective. The computational issues noted in Section 4.3.2 can be overcome with the additional assumption of a constant density over an interval prior to the monitoring time, that is, $f_1(t) = f$ for $t \in [C - W^*, C]$.

This data obtained from these diagnostic tests directly applies to the notation of Section 4.2 where, when two diagnostic tests are administered, a two-jump counting process or three-state progressive disease model is created. In this example, State 0 represents the uninfected state (or infected but not yet detectable by the sensitive test), State 1 represents the recently infected state (detectable by the sensitive diagnostic test but not yet by the less-sensitive test), and State 2 represents the non-recently infected state (detectable by both the sensitive and less-sensitive diagnostic tests). As before, the number of individuals in each state is denoted by n_{00}, n_{10}, n_{11}. Here the indices refer to the test results on the accurate and diluted diagnostic tests, respectively. For each index, a value of 1 indicates a positive result while a value of 0 indicates a negative result. In practice, the less-sensitive test is often given only when the sensitive test is positive. A negative result on the standard ELISA implies a negative result on the detuned ELISA. We assume no misclassification in the test results, an assumption which is discussed in further detail in Section 4.5. As in Balasubramanian and Lagakos (2010), we also assume that all individuals are tested at the same time and are the same age at this time of testing. This assumption in necessary in applications of infectious diseases such as HIV due to the age-period-cohort effect.

Recall Equation (4.6), the likelihood for a single monitoring time and a single waiting time, which we repeat here for convenience

$$(1 - F_1(C))^{n_{00}}(F_1(C) - F_1(C - W))^{n_{10}} F_1(C - W)^{n_{11}}.$$

We now add the additional assumption that the density $f_1(t) = f$ for t in the interval $[C - W^*, C]$, where $W^* \leq C$. This assumption that F_1 is locally linear implies that if you make such an assumption about F_1, then the only additional information needed to obtain the NPMLE of F_1 is the mean of the waiting times. To see this, assume W has finite support on $[0, W^*]$ and let $\overline{W}$ be the mean of the waiting times. The probability of being in each of the three states at the monitoring time C can be defined by

$$P(\text{in } S_0 \text{ at time C}) = 1 - F_1(C)$$

$$P(\text{in } S_1 \text{ at time C}) = f \int\limits_{C-W^*}^{C} (1 - G(C-t))dt = fE[W] = f\overline{W}$$

$$P(\text{in } S_2 \text{ at time C}) = 1 - (1 - F_1(C)) - f\overline{W} = (F_1(C) - f\overline{W}).$$

Incorporating this result into the likelihood above gives the following likelihood, for a single C with $f_1(t) = f$ in the interval $[C - W^*, C]$

$$(1 - F_1(C))^{n_{00}}(f\overline{W})^{n_{10}}(F_1(C) - f\overline{W})^{n_{11}}.$$

The maximum likelihood estimates of $F_1(C)$ and f, with $\overline{W}$ assumed known, can be obtained through standard numerical methods, yielding

$$\hat{F}_1(C) = 1 - \frac{n_{00}}{n} \quad \text{and} \quad \hat{f} = n_{10}/n\overline{W}.$$

The corresponding estimated HIV incidence rate at time C is

$$\hat{\lambda}(C) = \frac{f}{1 - \hat{F}_1(C)} = \frac{n_{10}}{n_{00}}\,\overline{W}.$$

The maximum likelihood estimate of $F_1(C - \overline{W})$ is then given by

$$\hat{F}_1(C - \overline{W}) = \frac{n_{11}}{n}.$$

Note that these estimators can be immediately obtained from the results of Section 4.3.1 by setting $W \equiv \overline{W}$, that is, by assuming the fixed known waiting time for all individuals is equal to the mean of the waiting time distribution. Therefore, under this assumption on the local behavior of F_1, the only additional information necessary to obtain the NPMLE of F_1 is the mean of waiting time distribution, a single fixed known value for all individuals, and hence the approach for a single waiting time can be used.

We are also interested in estimating the cumulative hazard of time to the first event in the recent past. We therefore consider estimation of this cumulative hazard over a short period, of length A, prior to the monitoring time, that is, over the interval $[C - A, C]$. Balasubramanian and Lagakos (2010) raise the question of selecting the optimal value of $\overline{W}$ with a view to precise estimation of the incidence rate. As a similar alternative, we consider the case where the waiting time between the two events is modifiable by design, and determine the optimal choice of $\overline{W}$ for estimation of the cumulative hazard of F_1 over $[C-A, C]$, for an appropriate choice of A. Through the use of simulated data, following the data structure of simultaneously administering both the standard and detuned ELISA to each individual, we examine determining this optimal waiting time for different underlying distributions, both where the constant density assumption holds and where it is violated. In either case, we estimate the cumulative hazard over the interval $[C-A, C]$ by

$$-\log\left(\frac{1-F(C)}{1-F(C-A)}\right). \tag{4.8}$$

First we consider the case where the assumption of a constant density truly exists for the underlying distribution F_1. In this case it is seen that the optimal choice of $\overline{W}$ is driven solely by the variance (Table 4.1). We then extend the analysis to the more realistic scenario where the density is no longer constant over the required interval. In this case, a bias-variance tradeoff must be considered (Table 4.2).

First consider simulated datasets of sample size $n = 500$, which are generated assuming the distribution function F_1 is Uniformly distributed with mean 1. The monitoring time is assumed fixed at $C = 1$ and values of $\overline{W}$ from 0 to 1 at equal increments of 0.1 are examined. Let the length of the interval of interest be fixed at $A = 0.5$. With this value of A and the fixed monitoring time at $C = 1$, the goal is to estimate the cumulative hazard of F_1 over the interval $[0.5, 1]$, the true value being 0.405. Table 4.1 shows the results of 1000 such simulations where the mean estimated cumulative hazard, the corresponding standard deviation and mean squared error (MSE) are presented for each assumed value of $\overline{W}$. The cumulative hazard is then estimated according to Equation (4.8). For this interval, $[0.5, 1]$, the corresponding estimate of the cumulative hazard is obtained through an estimate of

$$-\log\left(\frac{1-F(1)}{1-F(1-0.5)}\right).$$

When $\overline{W} \geq A$, it follows that $W^* \geq A$ and this calculation is straightforward given the assumption of a constant f over the interval $[C-W^*, C]$. However when $A > \overline{W}$, we make the additional assumption of a constant density over the interval $[C-A, C]$. This assumption is automatically satisfied if $W^* \geq A$, which is likely to be true in any application. As an estimate of $F(C-A)$ is not always directly observable, when $A \neq \overline{W}$, we estimate $(F-A)$ by simply extrapolating or interpolating the estimate of $F(C-\overline{W})$, using the constant density assumption.

$\overline{W}$	Mean Cumulative Hazard over interval [0.5, 1]	SD	MSE
0.1	0.406	0.072	0.00511
0.2	0.406	0.051	0.00262
0.3	0.406	0.044	0.00190
0.4	0.406	0.039	0.00151
0.5	0.407	0.037	0.00138
0.6	0.407	0.034	0.00117
0.7	0.407	0.032	0.00105
0.8	0.407	0.031	0.00098
0.9	0.407	0.030	0.00090
1.0	0.407	0.029	0.00086

Table 4.1: Simulated results ($n = 500$) for the mean estimated cumulative hazard over the interval $[0.5, 1]$ for a single monitoring time $C = 1$, with values of $\overline{W}$ from 0 to 1 at equal increments of 0.1. The true distribution F_1 is Uniformly distributed with mean 1. The true value of the cumulative hazard over the interval is therefore 0.405.

For a Uniform distribution, the density over all possible intervals will be constant. As can be seen from Table 4.1, when the true distribution F_1 is Uniformly distributed with mean 1, the optimal choice of $\overline{W}$ is to make $\overline{W}$ as large as possible while still satisfying the constraint that $\overline{W} \leq C$. Therefore, to estimate this cumulative hazard it is best to look beyond the interval of interest (in this case the interval of length $A = 0.5$), and interpolate to the required interval. In fact, this is true for all Uniform distributions regardless of the mean of the distribution. Assuming a Uniform distribution for F_1 implies that the locally linear assumption holds. This finding of the optimal $\overline{W}$ is as expected since extrapolation increases the standard deviation but interpolation reduces variability. As all values of $\overline{W}$ produce unbiased estimates of the cumulative hazard, the MSE is driven by the standard deviation. A Uniform distribution for F_1 follows the constant density assumption, and so making a linear extrapolation / interpolation does not introduce any bias. We therefore do not have to account for a bias-variance tradeoff, and can instead simply choose the optimal $\overline{W}$ as that with the lowest variance. Of course, the constant density assumption over $[C-W^*, C]$ is unlikely to be true for large values of W^*.

We therefore extend this scenario and consider finding the optimal $\overline{W}$ when the true underlying density is not constant over $[C - W^*, C]$. For large intervals the assumption of a constant density is not likely to hold true in many practical applications. Table 4.2 shows the results of similar simulations to those in Table 4.1 but where the true distribution F_1 now follows an Exponential distribution with mean 2. Again, sample sizes of $n = 500$ are considered and the monitoring time and interval length are fixed at $C = 1$ and $A = 0.5$, respectively. The cumulative hazard of F_1 is calculated in the same manner as that above (Equation (4.8)). The true value of this cumulative hazard over the interval $[0.5, 1]$ is 0.25.

$\overline{W}$	Mean Cumulative Hazard over interval [0.5, 1]	SD	MSE
0.1	0.225	0.053	0.00338
0.2	0.233	0.039	0.00181
0.3	0.238	0.033	0.00130
0.4	0.244	0.029	0.00090
0.5	0.250	0.027	0.00073
0.6	0.256	0.026	0.00069
0.7	0.263	0.024	0.00076
0.8	0.269	0.023	0.00091
0.9	0.275	0.023	0.00115
1.0	0.282	0.022	0.00150

Table 4.2: Simulated results ($n = 500$) for the mean estimated cumulative hazard over the interval $[0.5, 1]$ for a single monitoring time $C = 1$, with values of $\overline{W}$ from 0 to 1 at equal increments of 0.1. The true distribution F_1 is Exponentially distributed with mean 2. The true value of the cumulative hazard over the interval is therefore 0.25.

In Table 4.2 the distribution of F_1 does not follow the constant density assumption (F_1 is now Exponentially distributed with mean 2). In this case, although increasing the value of $\overline{W}$ decreases the variance, there is now a bias-variance tradeoff that needs to be considered. For this specific example where the constant density assumption fails, the optimal $\overline{W}$ for estimating the cumulative hazard of F_1 over the interval $[0.5, 1]$ is now to choose $\overline{W}$ just greater than A, the length of the interval. In Table 4.2 the optimal choice of $\overline{W}$, determined based on these 1000 simulations, is $\overline{W} = 0.6$. To verify this result, additional simulations were performed where the constant density assumption is also violated. Again the sample size, monitoring time, and interval length are fixed as before at $n = 500$, $C = 1$, and $A = 0.5$, respectively. Table 4.3 shows the results of one such additional simulation where F_1 is now assumed to follow a Weibull distribution with a hazard that doubles over the interval $[0.5, 1]$. The corresponding true cumulative hazard of F_1 over this interval is 0.75. Similar results to those in Table 4.2 are found under this scenario where the optimal choice of $\overline{W}$ is again to choose $\overline{W}$ just greater than A, specifically $\overline{W} = 0.6$.

$\overline{W}$	Mean Cumulative Hazard over interval [0.5, 1]	SD	MSE
0.1	0.715	0.090	0.00929
0.2	0.732	0.068	0.00490
0.3	0.745	0.060	0.00360
0.4	0.751	0.055	0.00298
0.5	0.750	0.052	0.00271
0.6	0.740	0.050	0.00258
0.7	0.722	0.049	0.00311
0.8	0.697	0.047	0.00499
0.9	0.662	0.044	0.00966
1.0	0.620	0.042	0.01863

Table 4.3: Simulated results ($n = 500$) for the mean estimated cumulative hazard over the interval $[0.5, 1]$ for a single monitoring time $C = 1$, with values of $\overline{W}$ from 0 to 1 at equal increments of 0.1. The true distribution F_1 is Weibull with a hazard that doubles over the interval $[0.5, 1]$. The true value of the cumulative hazard over the interval is therefore 0.75.

In practice, these simulations suggest that if you have an idea of a model, or a set of plausible models, you could perform such simulations to find the optimal $\overline{W}$. The optimal choice of $\overline{W}$ will depend on the estimate of interest and the corresponding interval of interest. As was seen here, if interest is on estimating the cumulative hazard over the recent past where the interval of interest is short, the density over this interval is unlikely to change, in which case the optimal choice of $\overline{W}$ is the maximum value of $\overline{W}$. However, if the interval of interest is longer, applicable to many practical examples, the density is likely to change within the interval. In this case, a choice of $\overline{W}$ is no longer to make $\overline{W}$ as large as possible, instead a value of $\overline{W}$ just beyond the length of the interval would be preferred.

4.5 Discussion

Throughout this paper we have considered a multistate model with exactly three states, specifically, a progressive three-state model. However, the methods and ideas presented here can easily be extended to allow for more distinct states, including Model (b) and Model (c) of Figure 4.1. Datta et al. (2009) consider a four state progressive model for an irreversible disease and also describe a more general acyclic multistate model.

We have discussed estimation of the cumulative hazard of the distribution of time to the first event in the recent past. The ideas are also easily extended to address the effect of a covariate on the distribution of time to the first event, and the recent cumulative hazard in particular. For the related data structure described in Section 4.1, consisting of a right censored observation of the final event with a current status observation of the intermediate event, regression models for the time until the first event, under proportional hazards, have been investigated by Dunson and Baird (2001) and Young et al. (2008).

Throughout this paper it is assumed that there is an ordering to the events, namely that the first event must always occur before the second event, thereby differing from bivariate current status data, discussed in Section 1.4.5. We assume an indicator that the first event has not yet occurred implies that the second event has also not yet occurred. However, if event occurrence is determined by a testing mechanism with imperfect sensitivity or specificity, a negative test result on the first event may not necessarily imply a true negative test result on the second event. The models discussed here can be extended to allow for misclassification (with known misclassification rates) using the ideas of Chapter 2 for misclassification of simple current status data. Furthermore, if interest is in estimating disease prevalence in the presence of misclassification, the group testing approach of Chapter 3 could also be applied to this multistate scenario.

In the application to simultaneous accurate and diluted HIV test data of Section 4.4, it is necessary to make the assumption that all individuals are the same age when tested, and that there exists only a single monitoring time C. This assumption is necessary in many practical examples as the underlying distribution of time to disease onset will not be the same for all ages as there is not only the affect of age, but there is also an underlying chronological time effect. In a single cross sectional sample, as considered in this chapter, age and time are confounded (Marschner, 1997).

Another issue is the potential for differential selection, as introduced in Section 1.10, across the various states. For the progressive three-state model considered here, differential selection would occur if the probability of being selected at a particular monitoring time differs from state to state. The presence of differential selection will lead to biased estimation of the cumulative hazard in the recent past. We assume there is no selection bias but more sophisticated models could be developed to allow for such a bias.

For the additional assumptions made in Section 4.4, that the density $f_1(t) = f$ for t in the interval $[C - W^*, C]$, Balasubramanian and Lagakos (2010) consider the association between a vector of covariates measured on each individual and the HIV incidence rate. They also address the issues of imperfect diagnostic tests and the impact of the risk of death on prevalence, under this assumption.

Chapter 5

Summary

This dissertation considers topics in current status data, some of which are well known and widely used, while others are more recent developments and have yet to receive much attention. Current status data is a type of survival data. Instead of the standard observation of failure times, with current status data the only available information on the failure time is whether or not the event has occurred before the examination time, generally referred to as the monitoring time. Although the current status data obtained appears somewhat limited, it can be seen throughout this dissertation that considerable attention has been given to analysis with this type of data. Many authors have also shown interest in current status data due to its non-standard asymptotic properties.

Chapter 1 gives an overview of current status data and some motivating examples to highlight some of the many areas in which current status data naturally occur in practice. The aim of Chapter 1 is to inform readers of the usefulness of current status data and to create an awareness of some of the available methods and resources for analyzing such data. Various areas of interest and importance in current status data are discussed in Chapter 1 including a description of the asymptotic properties, the correspondence between current status data and generalized linear models, the association with counting processes, and how to incorporate different sampling schemes, misclassification and competing risks into the current status data structure. Within each section, appropriate references are given, from which a more detailed description of the topic of interest can be obtained. This chapter may therefore prove beneficial to those interested in exploring current status data in more detail, or may simply make others more aware of the recent developments in the area.

The current status outcome is a binary variable simply indicating whether or not the event of interest has occurred before the monitoring time. The accuracy of this outcome and the impact of inaccurate outcomes is of great importance. Chapter 2 therefore considers current status data when the response variable is subject to (known) misclassification. In many epidemiological applications with imperfect diagnostic testing mechanisms all individuals examined may be subject to misclassification. It is seen in Chapter 2 how biased estimates of both the distribution function and the regression parameters are produced if this misclassification is ignored.

Many of the available testing mechanisms are costly and in many applications, such as in screening for HIV, resources are often limited. For this reason, the use of group testing has become a popular testing alternative. These diagnostic tests may again have inherent imperfect sensitivity or specificity. Chapter 3 therefore extends the results of Chapter 2 to consider group testing with current status data in the presence of misclassification. When misclassification exists, it is seen that group testing can be used for more efficient estimation of the prevalence of a rare disease than that of testing each individual sample separately. This approach also successfully reduces the number of necessary tests to be performed.

Group testing can be used to reduce the costs of testing individuals for a viral disease like HIV. With such an infectious disease, infected individuals are often asymptomatic, which results in many individuals not receiving treatment at an early stage, and others becoming infected unknowingly. Knowledge of the time at which individuals become infected is therefore important for both the individual and for health planning purposes. Chapter 4 considers estimation of the time until a positive result on a sensitive diagnostic test based on current status data, and specifically whether current status information on a subsequent event can be used to improve this estimate.

This dissertation introduces many concepts applicable to current status data and adds to the existing literature in the area by developing ideas not yet examined from a current status data perspective. There are many interesting open problems and possible extensions of this dissertation remaining for future work.

Bibliography

Anderson, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer: New York.

Anderson, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11:91–115.

Anttila, T., Saikku, P., Koskela, P., Bloigu, A., Dillner, J., Ikheimo, I., Jellum, E., Lehtinen, M., Lenner, P., Hakulinen, T., Nrvnen, A., Pukkala, E., Thoresen, S., Youngman, L., and Paavonen, J. (2001). Serotypes of chlamydia trachomatis and risk for development of cervical squamous cell carcinoma. *Journal of the American Medical Association*, 285(1):47–51.

Aschengrau, A. and Seage, G. R. (2003). *Essentials of Epidemiology in Public Health, Second Edition*. Jones & Bartlett Publishers.

ASHA (2011). American Social Health Association: HPV and cervical cancer prevention resource center. http://www.ashastd.org/hpv/hpv_learn.cfm.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26:641–647.

Balasubramanian, R. and Lagakos, S. W. (2010). Estimating HIV incidence based on combined prevalence testing. *Biometrics*, 66:1–10.

Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *Annals of Statistics*, 29:1699–1731.

Banerjee, M. and Wellner, J. A. (2005). Confidence intervals for current status data. *Scandinavian Journal of Statistics*, 32:405–424.

Barlow, R. E., Bartholomew, D. J., Brenner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley: New York.

Becker, N. G. (1989). *Analysis of Infectious Disease Data*. Chapman and Hall: New York.

Bickel, P. J., Gotze, F., and van Zwat, W. (1997). Resampling fewer than n observations: gains, losses and remedies for losses. *Statistica Sinica*, 7:1–31.

Bickel, P. J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 18:967–985.

Bilder, C. R., Tebbs, J. M., and Chen, P. (2010). Informative retesting. *Journal of the American Statistical Association*, 105(491):942–955.

Bosch, F. X., Manos, M. M., Muoz, N., Sherman, M., Jansen, A. M., Peto, J., Schiffman, M. H., Moreno, V., Kurman, R., and Shah, K. V. (1995). Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. international biological study on cervical cancer (IBSCC) study group. *Journal of the National Cancer Institute*, 87(11):796–802.

Bowman, J. A., Sanson-Fisher, R., and Redman, S. (1997). The accuracy of self-reported Pap smear utilisation. *Social Science and Medicine*, 44(7):969–976.

Brookmeyer, R. (1999). Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics*, 55:608–612.

Burr, D. and Gomatam, S. (2002). On nonparametric regression for current status data. National Institute of Statistical Sciences, Technical Report Numbeer 127.

Campanelli, P. C., Salo, M. T., Schwede, L., and Jackson, B. (2007). The accuracy of self-reports: Some preliminary findings from interviewing homeless persons. U.S. Census Bureau, Statistical Research Division, Research Report Series, Survey Methodology 2007-7.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91(433):242–250.

CDC (1999). Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome. *Mortality and Morbidity Weekly Report (MMWR)*, 48(RR-13):1–28.

CDC (2000). Centers for Disease Control and Prevention: Tracking the hidden epidemic: Trends in STDs in the United States. Available at www.cdc.gov/std/stats98/STD_Trends.pdf.

CDC (2008). HIV Prevalence estimates - United States, 2006. *Mortality and Morbidity Weekly Report (MMWR)*, 57(39):1073–1076.

CDC (2011a). Centers for Disease Control and Prevention: HIV/AIDS. Available at www.cdc.gov/hiv/.

CDC (2011b). Centers for Disease Control and Prevention: Human Papillomavirus (HPV). http://www.cdc.gov/std/HPV/.

CDC (2011c). National Health Care Surveys. Available at http://www.cdc.gov/nchs/dhcs/about_dhcs.htm.

Chen, M., Ibrahim, J. G., and Sinha, D. (2004). A new joint model for longitudinal and survival data with a cure fraction. *Journal of Multivariate Analysis*, 91:18–34.

Cheng, G. and Wang, X. (2011). Semiparameric additive transformation model under current status data. Cornell University Library. Available at http://arxiv.org/abs/1105.1304v1.

Cook, R. J. and Tolusso, D. (2009). Second-order estimating equations for the analysis of clustered current status data. *Biostatistics*, 10(4):756–772.

Datta, S., Lan, L., and Sundaram, R. (2009). Nonparametric estimation of waiting time distributions in a Markov model based on current status data. *Journal of Statistical Planning and Inference*, 139:2885–2897.

Datta, S., Satten, G. A., and Sundaram, R. (2000). Nonparametric estimation for a three-stage irreversible illness-death model. *Biometrics*, 56:841–847.

Datta, S. and Sundaram, R. (2006). Nonparametric estimation of stage occupation probabilities in a multistage model with current status data. *Biometrics*, 62:1–38.

de Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24.

DeGruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data with applications to AIDS. *Biometrics*, 45:1–11.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:829–837.

DHS (2011). MEASURE DHS: Demographic and Health Surveys. Available at http://www.measuredhs.com/data.

Diamond, I. D. and McDonald, J. W. (1991). The analysis of current status data. In Demographic Applications of Event History Analysis. Oxford UK: Oxford University Press.

Diamond, I. D., McDonald, J. W., and Shah, I. H. (1986). Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography*, 23:607–620.

Ding, A. A. and Wang, W. (2004). Teating independece for bivariate current status data. *Journal of the American Statistical Association*, 99(465):145–155.

Dinse, G. E. and Lagakos, S. W. (1982). Nonparametric estimation of lifetime and disease onset distributions from incomplete observations. *Biometrics*, 38:921–932.

Doksum, K. A. and Gasko, M. (1990). On a correspondence between models in binary regression and in survival analysis. *International Statistical Review*, 58:243–252.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14:436–440.

Dunson, D. B. and Baird, D. D. (2001). A flexible parametric model for combining current status and age at first diagnosis data. *Biometrics*, 57:396–403.

Dunson, D. B. and Dinse, G. E. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics*, 58(1):79–88.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Eng, T. and Butler, W. (1997). The hidden epidemic: Confronting sexually transmitted diseases. Washington, DC: National Academy Press.

Fang, H. B., Li, G., and Sun, J. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, 32:59–75.

Fleming, T. R. and Harrington, D. P. (2005). *Counting Processes and Survival Analysis.* Wiley Series in Probability and Statistics.

Fowles, J. B., Rosheim, K., Fowler, E. J., Craft, C., and Arrichiello, L. (1999). The validity of self-reported diabetes quality of care measures. *International Journal for Quality in Health Care*, 11:407–412.

Galbraith, S., Daniel, J. A., and Vissel, B. (2010). A study of clustered data and approaches to its analysis. *The Journal of Neuroscience*, 30(32):10601–10608.

Gart, J. J., Krewski, D., Lee, P. N., Tarone, R., and Wahrendorf, J. (1986). Statistical methods in cancer research, Volume III, the design and analysis of longterm animal experiments. IARC Scientific Publications No. 79 Lyon: International Agency for Research on Cancer.

Gibson, D. R., Lovelle-Drache, J., Young, M., Hudes, E. S., and Sorensen, J. L. (1999). Effectiveness of brief counseling in reducing HIV risk behavior in injecting drug users: final results of randomized trials of counseling with and without HIV testing. *AIDS and Behavior*, 3:3–12.

Graff, L. E. and Roeloffs, R. (1972). Group testing in the presence of test errors: an extension of the Dorfman procedure. *Technometrics*, 14:113–122.

Groeneboom, P., Maathuis, M. H., and Wellner, J. A. (2008). Current status data with competing risks: consistency and rates of convergence of the MLE. *Annals of Statistics*, 36:1031–1063.

Groeneboom, P. and Wellner, J. A. (1992). *Nonparametric Maximum Likelihood Estimators for Interval Censoring and Denconvolution*. Boston: Birkhauser-Boston.

Grummer-Strawn, L. M. (1993). Regression analysis of current status data: an application to breast-feeding. *Journal of the American Statistical Association*, 88:758–765.

Hall, H. I., Song, R., Rhodes, P., Prejean, P., An, Q., Lee, L. M., Karon, J., Brookmeyer, R., Kaplan, E. H., McKenna, M. T., and Janssen, R. S. (2008). Estimation of HIV incidence in the United States. *JAMA*, 300(5):520–529.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimation. *Journal of Econometrics*, 35:303–316.

Hardin, J. W., Schmiediche, H., and Carroll, R. J. (2003). The simulation extrapolation method for fitting generalized linear models with additive measurement error. *The Stata Journal*, 3(4):1–12.

Hens, N., Weinke, A., Aerts, M., and Molenberghs, G. (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine*, 28:2785–2800.

Ho, G. Y., Bierman, R., Beardsley, L., Chang, C. J., and Burk, R. D. (1998). Natural history of cervicovaginal papillomavirus infection in young women. *New England Journal of Medicine*, 338(7):423–428.

Honda, T. (2004). Nonparametric regression with current status data. *Annals of the Institute of Statistical Mathematics*, 56(1):49–72.

Hu, P., Tsiatis, A. A., and Davidson, M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics*, 54:1407–1419.

Huang, J. (1996). Efficient estimation for the cox model with interval censoring. *Annals of Statistics*, 24:540–568.

Huang, J. and Wellner, J. A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I. *Statistica Neerlandica*, 49:153–163.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis, D. Y. Lin and T. R. Fleming (eds), 123-169. New York: Springer-Verlag.

Hudgens, M. G., Satten, G. A., and Jr., I. M. L. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics*, 57:74–80.

Hughes-Oliver, J. M. and Swallow, W. H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association*, 89:982–993.

Janssen, R. S., Satten, G. A., Stramer, S. L., Rawal, B. D., OBrien, T. R., Weiblen, B. J., Hecht, F. M., Jack, N., Cleghorn, J., Kahn, J. O., Chesney, M. A., and Busch, M. P. (1998). New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *Journal of the American Medical Association*, 280:42–48.

Jay, N. and Moscicki, A.-B. (2000). Human papilloma virus infection in women. Marlene Goldman and Maureen Hatch, eds. Women and Health. San Diego, CA: Academic Press.

Jewell, N. P. and Kalbfleisch, J. D. (2004). Maximum likelihood estimation of ordered multinomial parameters. *Biostatistics*, 5(2):291–306.

Jewell, N. P., Malani, H., and Vittinghoff, E. (1994). Nonparametric estimation for a form of doubly censored data with application to two problems in AIDS. *Journal of the American Statistical Association*, 89:7–18.

Jewell, N. P. and Shiboski, S. C. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics*, 46:1133–1150.

Jewell, N. P. and van der Laan, M. J. (1995). Generalizations of current status data with applications. *Lifetime Data Analysis*, 1:101–109.

Jewell, N. P. and van der Laan, M. J. (2004a). Case-control current status data. *Biometrika*, 91(3):529–541.

Jewell, N. P. and van der Laan, M. J. (2004b). Current status data: review, recent developments and open problems. In Advances in Survival Analysis, Handbook in Statistics #23: 625-642, Amsterdam: Elsevier.

Jewell, N. P., van der Laan, M. J., and Henneman, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika*, 90:183–197.

Jewell, N. P., van der Laan, M. J., and Lei, X. (2005). Bivariate current status data with univariate monitoring times. *Biometrika*, 92(4):847–862.

Jewell, N. P., van der Laan, M. J., and Shiboski, S. (2006). Choice of monitoring mechanism for optimal nonparametric functional estimation for binary data. *The International Journal of Biostatistics*, 2(1):Article 7.

Kalbfleish, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* 2nd Ed. New Jersey: Wiley.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society A.*, 154:371–412.

Keiding, N., Begtrup, K., Scheike, T. H., and Hasibeder, G. (1996). Estimation from current status data in continuous time. *Lifetime Data Analysis*, 2:119–129.

Kline, R. L., Brothers, T. A., Brookmeyer, R., Zeger, S., and Quinn, T. C. (1989). Evaluation of human immunodeficiency virus (HIV) sero-prevalence in population surveys using pooled sera. *Journal of Clinical Microbiology*, 27:1449–1455.

Koutsky, L. (1997). Epidemiology of genital human papillomavirus infection. *American Journal of Medicine*, 102(5A):3–8.

Koutsky, L. and Kiviat, N. (1999). Genital Human Papillomavirus. King Holmes et al. eds. Sexually Transmitted Diseases, 3rd ed. New York: McGraw-Hill.

Krailo, M. D. and Pike, M. C. (1983). Estimation of the distribution of age at natural menopause from prevalence data. *American Journal of Epidemiology*, 117:356–361.

Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*, 62:85–96.

Lam, K. and Xue, H. (2005). A semiparametric regression cure model with current status data. *Biometrika*, 92:573–586.

Lan, L. and Datta, S. (2008). Non-parametric estimation of state occupation, entry and exit times with multistate current status data. *Statistical Methods in Medical Research*, 19:147–165.

Lequin, R. M. (2005). Enzyme immunoassay (eia)/enzyme-linked immunosorbent assay (elisa). *Clinical Chemistry*, 51(12):2415–2418.

Lesthaeghe, R. J. and Page, H. J. (1980). The post-partum non-susceptible period: development and application of model schedules. *Population Studies*, 34:143–169.

Li, Z. and Nan, B. (2011). Relative risk regression for current status data in case-cohort studies. *The Canadian Journal of Statistics*, 9999:1–21.

Lim, L. L. Y., Seubsman, S., and Sleigh, A. (2009). Validity of self-reported weight, height, and body mass index among university students in thailand: Implications for population studies of obesity in developing countries. *Population Health Metrics*, 7:7–15.

Lin, D. Y., Oakes, D., and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, 85:289–298.

Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88:1341–1349.

Litvak, E., Tu, X. M., and Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association*, 89:424–434.

Liu, A., Liu, C., Zhang, Z., and Albert, P. S. (2011). Optimality of group testing in the presence of misclassification. *Biometrika*, Under:Review.

Liu, H. and Shen, Y. (2009). A semiparametric regression cure model for interval-censored data. *Journal of the American Statistical Association*, 104(487):1168–1178.

Lu, W. and Ying, Z. L. (2004). On semiparameteric transformation cure model. *Biometrika*, 91:331–343.

Ma, S. (2007). Additive risk model with case-cohort sampled current status data. *Statistical Papers*, 48:595–608.

Ma, S. (2009). Cure model with current status data. *Statistica Sinica*, 19:233–249.

Ma, S. (2011). Additive risk model for current staus data with a cured subgroup. *Annals of the Institute of Statistical Mathematics*, 63:117–134.

Ma, S. and Kosorok, M. R. (2005). Penalized log-likelihood estimation for partly linear transformation models with current status data. *Annals of Statistics*, 33:2256–2290.

Maathuis, M. H. (2011). Nonparametric inference for competing risks current status data with continuous, discrete or grouped observation times. *Biometrika*, 98(2):325–340.

MacMahon, B. and Worcester, J. (1966). Age at menopause, United States 1960 to 1962. National Center for Health Statistics; Vital and Health Statistics, Series 11; Data from the National Health Survey, No. 19. Washington, DC: DHEW Publication no. (HSM) 66-1000.

Maller, R. A. and Zhou, X. (2001). *Survival Analysis with Long term Survivors*. New York. Wiley.

Marschner, I. C. (1997). A method for assessing age-time disease incidence using serial prevalence data. *Biometrics*, 53:1384–1398.

Martinussen, T. and Scheike, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika*, 89:649–658.

McDermott-Webster, M. (1999). The HPV epidemic. *American Journal of Nursing*, 99:24L–24N.

Meira-Machado, L., Cadarso-Suarez, C., and Una-Alvarez, J. (2008). Inference in multi-state survival data. Proceedings of the 13th International Conference on Applied Mathematics.

Meira-Machado, L., de Una-Alverez, J., Cadorso-Suarez, C., and Anderson, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, 18:195–222.

Miettinen, O. S. (1982). Design options in epidemiologic research. An update. *Scandinavian Journal of Work and Environmental Health*, 8(1):7–14.

Moscicki, A.-B., Shiboski, S., Broering, J., Powell, K., Clayton, L., Jay, N., Darragh, T. M., Brescia, R., Kanowitz, S., Miller, S. B., Stone, J., Hanson, E., and Palefsky, J. (1998). The natural history of human papillomavirus infection as measured by repeated dna testing in adolescent and young women. *Journal of Pediatrics*, 132:277–284.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86:843–855.

NIAID (2011). National Institute of Allergies and Infectious Diseases: HIV/AIDS. http://www.niaid.nih.gov/TOPICS/HIVAIDS/.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling.* New York: Springer.

Qi, L., Wang, C. Y., and Prentice, R. L. (2005). Weighted estimator for proportional hazards with missing covariate regression. *Journal of the American Statistical Association*, 100:1251–1263.

Quinn, T. C., Wawer, M. J., Sewankambo, N., Serwadda, D., Li, C., Wabwire-Mangen, F., Meehan, M. O., Lutalo, T., and Gray, R. H. (2000). Viral load and heterosexual transmission of human immunodeficiency virus type 1. *New England Journal of Medicine*, 342(13):921–929.

Rabinowitz, D. and Jewell, N. P. (1996). Regression with doubly censored current status data. *Journal of the Royal Statistical Society B*, 58(3):541–550.

Rhodes, F. and Malotte, C. K. (1996). HIV risk interventions for active drug users. S. Oskamp, S. Thompson, eds. Understanding HIV risk behavior: safer sex and drug use. Thousand Oaks, CA: Sage Publications.

Rietmeijer, C. A., Kane, M. S., Simons, P. Z., Corby, N. H., Wolitski, R. J., Higgins, D. L., Judson, F. N., and Cohn, D. L. (1996). Increasing the use of bleach and condoms among injecting drug users in denver: outcomes of a targeted, community level HIV prevention program. *AIDS*, 10(3):291–298.

Rossini, A. J. and Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, 91:713–721.

Sal y Rosas, V. G. and Hughes, J. P. (2011). Nonparametric and semiparametric analysis of current status data subject to outcome misclassificaton. *Statistical Communications in Infectious Diseases*, 3(1):Article 7.

Santana, L. (2009). Contributions to the m-out-of-n bootstrap. Available at http://hdl.handle.net/10394/4906.

Sato, T. (1992). Maximum likelhood estimation of the risk ratio in case-cohort studies. *Biometrics*, 48(1215–1221):4.

Schiller, J. T., Day, P. M., and Kines, R. C. (2010). Current understanding of the mechanism of HPV infection. *Gynecologic Oncology*, 118:S12–S17.

Sen, B., Banerjee, M., and Woodroofe, M. (2010). Inconsistency of bootstrap: the Grenander estimator. *Annals of Statistics*, 38(4):1953–1977.

Sepkowitz, K. A. (2001). AIDS - the first 20 years. *New England Journal of Medicine*, 344(23):1764–1772.

Shen, X. (2011). Linear regression with current status data. *Journal of the American Statistical Association*, 95(451):842–852.

Shiboski, S. C. (1998a). Generalized additive models for current status data. *Lifetime Data Analysis*, 4:29–50.

Shiboski, S. C. (1998b). Partner studies. In The Encyclopedia of Biostatistics, P. Armitage and T. Colton (eds.).

Shiboski, S. C. and Jewell, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association*, 87:360–372.

Short, M. E., Goetzel, R. Z., Pei, X., Tabrizi, M. J., Ozminkowski, R. J., Gibson, T. B., DeJoy, D. M., and Wilson, M. G. (2009). How accurate are self-reports? an analysis of self-reported healthcare utilization and absence when compared to administrative data. *Journal of Occupational and Environmental Medicine*, 51(7):786–796.

Song, X. and Huang, Y. (2005). Multiple augmentation for interval-censored data with measurement error. *Statistics in Medicine*, 27:3178–3190.

Studts, J. L., Ghate, S. R., Gill, J. L., Studts, C. R., Barnes, C. N., LaJoie, A. S., Andrykowski, M. A., and LaRocca, R. V. (2006). Validity of self-reported smoking status among participants in a lung cancer screening trial. *Cancer Epidemiology, Biomarkers and Prevention*, 15:1825–1828.

Sun, J. (2006). *The Statistical Analysis of Interval Censored Failure Time Data.* Springer: New York.

Sun, J. and Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. *Journal of the American Statistical Association*, 88(424):1449–1454.

Sun, J. and Sun, L. (2005). Semiparametric linear transformation models for current status data. *The Canadian Journal of Statistics*, 33:85–96.

Sutradhar, R., Barbera, L., Seow, H., Howell, D., Husain, A., and Dudgeon, D. (2010). Multistate analysis of interval-censored longitudinal data: Application to a cohort study on performance status among patients diagnosed with cancer. *American Journal of Epidemiology*, 174(4):468–475.

Thompson, L. A. and Chhikara, R. S. (2003). A bayesian cure rate model for repeated measurements and interval censoring. Proceedings of JSM 2003.

Tian, L. and Cai, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika*, 93:329–342.

Tong, X., Zhu, C., and Sun, J. (2007). Regression analysis of two-sample current status data, with applications to tumorigenicity experiments. *The Canadian Journal of Statistics*, 35(4):575–584.

Tsiatis, A. A. and Davidson, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88:447–458.

Tu, X. M., Litvak, E., and Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating the prevalence of a rare disease: application to HIV screening. *Biometrika*, 82(2):287–297.

van der Laan, M. J., Bickel, P., and Jewell, N. P. (1997a). Singly and doubly censored current status data: estimation, asymptotics and regression. *Scandinavian Journal of Statistics*, 24(3):289–307.

van der Laan, M. J. and Jewell, N. P. (2001). The NPMLE for doubly censored current status data. *Scandinavian Journal of Statistics*, 28:537–547.

van der Laan, M. J. and Jewell, N. P. (2003). Current status and right-censored data structures when observing a marker at the censoring time. *The Annals of Statistics*, 31:512–535.

van der Laan, M. J., Jewell, N. P., and Peterson, D. R. (1997b). Efficient estimation of the lifetime and disease onset distribution. *Biometrika*, 84:539–554.

Vaupel, J., Manton, K., and Stallard, E. (1979). The impact of heterogeneity in individual frailty in the dynamics of mortality. *Demography*, 16(3):439–454.

Verdon, M. (1997). Issues in the management of human papillomavirus genital disease. *American Family Physician*, 55:1813–1820.

Villa, L. L. and Denny, L. (2006). Methods for detection of HPV infection and its clinical utility. *International Journal of Gynecology and Obstetrics*, 94 (Supplement 1):S71–S80.

Wang, L., Sun, J., and Tong, X. (2008). Efficient estimation for the proportional hazards model with bivariate current status data. *Lifetime Data Analysis*, 14:134–153.

Wang, R. and Lagakos, S. W. (2010). Augmented cross-sectional prevalence testing for estimating hiv incidence. *Biometrics*, 66:864–874.

Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika*, 87(4):879–893.

Wen, C.-C. and Chen, Y.-H. (2011). Nonparametric maximum likelihood analysis of clustered current status data with the gamma-frailty cox model. *Computational Statistics and Data Analysis*, 55:1053–1060.

Wen, C.-C., Huang, S. Y. H., and Chen, Y.-H. (2011). Cox regression for current status data with mismeasured covariates. *Canadian Journal of Statistics*, 39(1):73–88.

Wen, C.-C. and Lin, C.-T. (2011). Analysis of current status data with missing covariates. *Biometrics*, 69(3):760–769.

WHO (2007). World Health Organization: 2007 AIDS epidemic update. Available at data.unaids.org/pub/epislides/2007/2007_epiupdate_en.pdf.

Xie, M., Tatsuoka, K., Sacks, J., and Young, S. S. (2001). Group testing with blockers and synergism. *Journal of the American Statistical Association*, 96:92–102.

Xue, H., Larn, K. F., and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association*, 99:346–356.

Ying, G.-S. and Liu, C. (2006). Statistical analysis of clustered data using SAS system. Northeast SAS Users Group (NESUG). Available at
www.nesug.org/proceedings/nesug06/an/da01.pdf.

Young, J. G., Jewell, N. P., and Samuels, S. J. (2008). Regression analysis of a disease onset distribution using diagnosis data. *Biometrics*, 64:20–28.

Zagurski, K. (2006). Douglas County rates B+ on meetings in its health goals, but Dr Adi Pour says there is a lot of work to be done on reducing STDs. Omaha World Herald, Feb. 2, p. 08B.

Zhang, Z. and Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, 91:53–70.

Zhang, Z., Sun, J., and Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine*, 24:1300–1407.

Zhou, H. and Pepe, M. S. (1995). Auxiliary covariate data in failure time regression analysis. *Biometrika*, 82:139–149.

Appendices

Appendix A

Additional Data Applications

As described in Chapter 2, there are many applications in which the current status outcomes could be measured with error. Misclassification can occur due to laboratory error, imperfect testing mechanisms, misreporting, and other factors. The illustration in Chapter 2 focuses specifically on an application to first HPV infection among women in San Francisco. The assumed classification rates of $\alpha = 0.73$ and $\beta = 0.9$ used in this example arise from an imperfect testing mechanism, combined with the potential for recovery due to the nature of HPV infection. Misclassification that varies over time is also addressed (Section 2.2.3) where the probability of a false negative decreases when an individual is tested far from the time of the true event, the time of first HPV infection. However, these classification rates are specific to this particular case of estimating the time until first HPV infection, and are not necessarily appropriate for all applications. This appendix considers misclassification of current status data for different examples, using different levels of misclassification than those presented in Chapter 2. The results of similar analyses to those performed for the HPV application are presented in the following sections for two additional datasets. The results of this appendix can therefore be combined with those of Chapter 2 to give an overall understanding of the impact of failing to consider misclassification when estimating the time until a specific event of interest, using current status data.

A.1 Datasets

The two additional datasets we consider in this appendix are both obtained through a survey instrument based on self reporting. Self reporting is one of the most widely used methods of data collection for information regarding the health status of individuals. This approach has been used to assess an extensive range of health behaviors, including estimating the prevalence of certain risk factors, the utilization of available measures of preventive care, and the use of mental healthcare services. The accuracy of self reports have received considerable attention in a wide variety of fields, for example, when examining healthcare utilization (Short et al., 2009), identifying homeless persons (Campanelli et al., 2007), assessing the duration of work disability (Campanelli et al., 2007), and studying obesity (Lim et al., 2009). The reliability of self reports have also been considered in applications to examine various aspects of compliance and utilization of various other public health areas of interest, such as, diabetes (Fowles et al., 1999), lung cancer (Studts et al., 2006), and pap smears (Bowman et al., 1997), amongst others. Through the datasets considered in this appendix, we examine estimating the time until weaning, and the time until natural menopause. Although the levels of misreporting vary depending on the type of survey being carried out, the questions of interest, and whether there is social pressure to give a specific response, it is widely agreed that self reports are subject to misclassification, some of which may simply be due to lack of knowledge of the respondent. Therefore, both datasets considered below are immediately subject to potential misclassification. Different sensitivity and specificity rates, some of which may be due to respondents having reasons for hiding or distorting their true answers, are examined for each application.

Breastfeeding Example:

The Demographic and Health Surveys (DHS) are a series of studies conducted in developing countries to study population and health. The DHS program has been implemented in over 90 countries across Africa, Asia, Latin America and the Caribbean since it commenced in 1984. The health information obtained through these surveys is commonly used to help plan, monitor, and evaluate population, health and nutrition programs in these developing countries. As this initiative is publicly funded, the data collected is made available through the DHS website (DHS, 2011). Such datasets have thus been widely considered and analyzed from various approaches. The first additional dataset we consider was obtained from a DHS survey carried out in Indonesia. The data we consider consists of 8041 women between the ages of 15 and 49. For each woman in the study, a binary variable indicating whether their child has been weaned or not, and the current age of the child (in months) are observed, thus giving us current status data. Covariates describing the residential location and educational level of each mother are also available, along with a variety of other covariates. This dataset, along with DHS surveys for other locations are considered by Shiboski (1998a) and Grummer-Strawn (1993), where additional information on our dataset is provided.

Menopause Example:

The National Health Care Surveys are a set of surveys carried out in the United States by the Center for Disease Control and Prevention (CDC). These nationally representative surveys are designed to answer key questions of interest to healthcare policy makers, public health professionals, and researchers. Each survey collects core information which remains stable over time, as well as other key elements, including the factors that influence the use of healthcare resources, the quality of healthcare, and disparities in healthcare services provided to population subgroups. More information on these surveys can be found on the CDC website (CDC, 2011c). The second dataset we consider in this appendix is based on data obtained from the National Center for Health Statistics' Health Examination Survey. This dataset has previously been analyzed by several authors, where more information on the dataset can also be obtained (Jewell et al., 2003, Krailo and Pike, 1983, MacMahon and Worcester, 1966). In this survey the menopausal history of female respondents is examined where the age and menopausal status at the time of the survey are observed for each woman. For those who have experienced menopause, whether menopause was natural or operative is also recorded. Jewell et al. (2003) analyze this data ($n = 2423$) from a current status data perspective with two competing risks for the menopausal outcome, see Section 1.7 for more information on competing risks current status data. Here we are interested in estimating the time until natural menopause and therefore exclude the data for those who experienced operative menopause, thus reducing our dataset to $n = 2076$.

The remainder of this appendix is organized in a similar manner to Chapter 2. The sections, figures and tables presented here are in the same general order as the application to HPV infection (Section 2.2.2) to allow for straightforward comparison of results. For each table and figure, the results of the breastfeeding example are presented first, followed by those for the menopause example. Within each section, the level of misclassification assumed for each dataset is clearly identified.

A.2 Nonparametric Estimation of a Single Distribution Function

Recall the definition of the classification rates, α and β, given in Chapter 2

$$P(\Delta = 1|Y = 1) = \alpha \quad P(\Delta = 0|Y = 0) = \beta.$$

For the DHS dataset, we are primarily interested in estimating the time until weaning in Indonesia. In this example, we assume constant known classification rates of $\alpha = 0.9$ and $\beta = 1$. These rates imply that all observations indicating a child has been weaned at the time of the survey are correct ($\beta = 1$). However, we assume there is the potential for imperfect classification ($\alpha = 0.9$) when observing a child is still being breastfed at the time of the survey. These rates seem reasonable as there exists a social desirability bias where breastfeeding is considered socially desirable. On the other hand, we see no reason for a mother to misreport and say her child has been weaned, if she is in fact still breastfeeding. We also recognize that the desire to misreport may be greater if you have recently ceased breastfeeding, a concept readdressed in Section A.2.2.

For the menopause data obtained from the National Health Care Surveys, we assume constant known classification rates of $\alpha = 0.95$ and $\beta = 0.9$. In this instance, as the menopausal status of a woman is measured through self-reporting, although there may not be social pressure to misreport, the possibility of imperfect sensitivity or specificity still exists. For an silent event like menopause, the exact timing of this event is not immediately apparent, therefore leading to the unintentional misreporting of information. As illustrated in Section A.2.2, if a woman is examined close to the time at which menopause occurs, there is increased potential for misclassification of the menopausal status.

Figure A.1, similar to Figure 2.2, shows the estimated cumulative distribution functions for both the breastfeeding and menopause examples, under no misclassification and constant misclassification at the specified rates. Figure A.1(a) displays estimates of the time until weaning, where the time scale is measured by age, in months. The unconstrained NPMLE and the NPMLE adjusted for constant misclassification at the assumed rates of $\alpha = 0.9$ and $\beta = 1$ are shown. Figure A.1(b) considers the time until natural menopause with age now measured in years, and shows the unconstrained NPMLE and the NPMLE adjusted for constant misclassification with known classification rates of $\alpha = 0.95$ and $\beta = 0.9$.

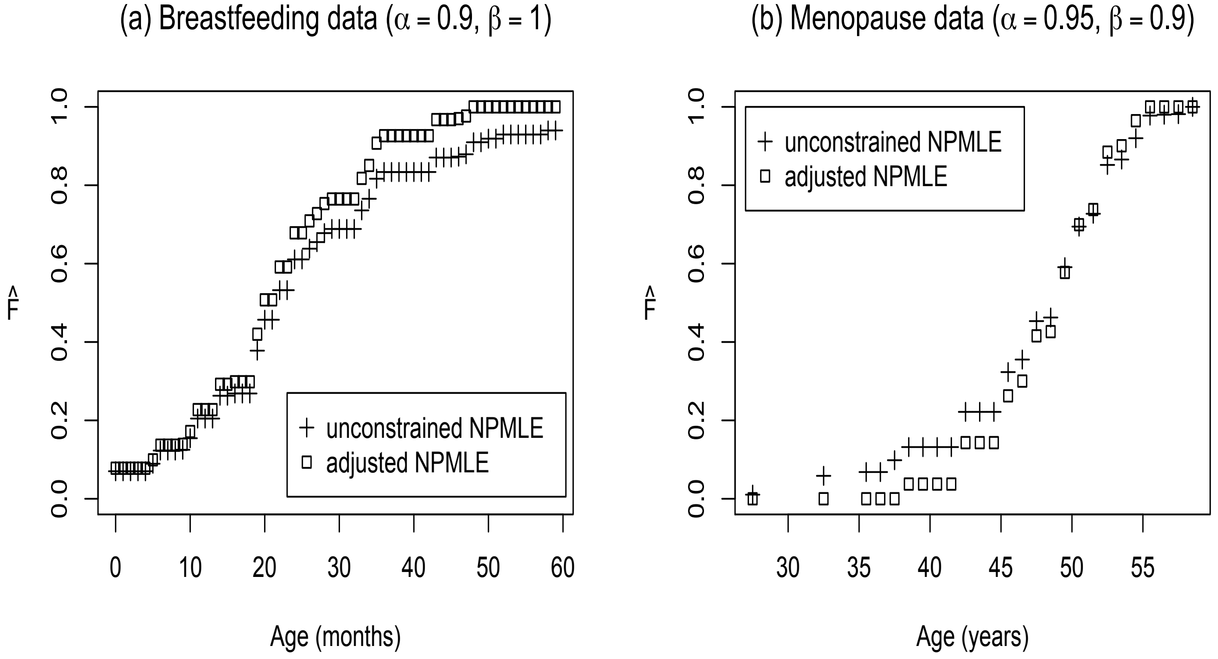

Figure A.1: Estimated cumulative distribution functions for (a) the breastfeeding example and (b) the menopause example. Both the unconstrained NPMLE obtained through the pool-adjacent violators algorithm and the proposed adjusted NPMLE are presented for each example.

When $\beta = 1$, which occurs only when there is perfect specificity, the two curves (the adjusted and unconstrained NPMLEs) will intersect at $\widehat{F} = 0$. Note that perfect specificity is assumed in Figure A.1(a). In this plot the curves are not seen to intersect at $\widehat{F} = 0$ as the value of $\widehat{F}$ at the first monitoring time is already greater than zero. As shown in Figure A.1(a), with perfect specificity, the adjusted NPMLE will always be greater than (or equal to, at $\widehat{F} = 0$) the unconstrained NPMLE. This is true for all cases where there is perfect specificity and imperfect sensitivity, regardless of the level of imperfect sensitivity. Of course, the closer the sensitivity rates are to perfect, the closer the two curves will be to each other, at each of the monitoring times. The reverse will be seen if there is perfect sensitivity and imperfect specificity, where the adjusted NPMLE will always be less than or equal to the unconstrained NPMLE, with the curves then intersecting at $\widehat{F} = 1$.

Recall, from the hypothetical example of Section 2.2.2, that when the classification rates are equal, the two curves will cross at $\widehat{F} = 0.5$, the estimated median time to occurrence. This result holds for all equal classification rates of $\alpha = \beta > 0.5$. At monitoring times below the point of intersection, the adjusted NPMLE is shifted downwards from the unconstrained estimator, and similarly shifted upwards at monitoring times above the point of intersection. Figure A.1(b) assumes unequal classification rates with both imperfect sensitivity and specificity. With the assumed rates of $\alpha = 0.95$ and $\beta = 0.9$ used in the menopause example, the two estimators will cross at a point greater than the median time to occurrence ($\widehat{F} = 0.667$) as the sensitivity is greater than the specificity. The reverse was seen in the application considering the time until first HPV infection given in Section 2.2.2 were the specificity was greater than the sensitivity.

A.2.1 Confidence Intervals

For reasons described in Section 2.2.1, the standard bootstrap yields inconsistent estimates of the pointwise confidence intervals with current status data. As an alternative, we use the m out of n bootstrap of Section 2.2.1, on which more details are given in Banerjee and Wellner (2005), for confidence interval estimation. Again, we consider symmetric confidence intervals as they often perform better in finite samples. Table A.1 and Table A.2 show the results of these symmetric confidence intervals for both the breastfeeding and menopause examples, respectively.

The 95% symmetric confidence intervals, presented below, were calculated for the adjusted NPMLE, using the appropriate levels of classification for each application. As with the results of Section 2.2.2, we are not interested in estimating the optimal block size m for use with the m out of n bootstrap. Instead, the results are given for three monitoring times at various values of m. The block sizes again represent the block sizes used by Politis et al. (1999), that is, values of m defined by $\{n^{1/3}, n^{1/2}, n^{2/3}, n^{3/4}, n^{0.8}, n^{0.9}\}$.

t_0	16 months	22 months	31 months
$\hat{F}(t_0)$	0.299	0.591	0.765
$m = 20$ $(n^{1/3})$	[0.214 0.384]	[0.506 0.676]	[0.695 0.835]
$m = 90$ $(n^{1/2})$	[0.219 0.379]	[0.496 0.686]	[0.710 0.820]
$m = 401$ $(n^{2/3})$	[0.229 0.369]	[0.501 0.681]	[0.700 0.830]
$m = 849$ $(n^{3/4})$	[0.234 0.346]	[0.486 0.696]	[0.705 0.825]
$m = 1331$ $(n^{0.8})$	[0.239 0.359]	[0.501 0.681]	[0.700 0.830]
$m = 3271$ $(n^{0.9})$	[0.249 0.349]	[0.511 0.671]	[0.715 0.815]

Table A.1: Breastfeeding example: estimation of the time until weaning. Confidence interval estimation for the adjusted ($\alpha = 0.9$, $\beta = 1$) NPMLE at three monitoring times obtained using the m out of n bootstrap for various values of m ranging from 20 to 3271.

Table A.1 shows the 95% symmetric confidence intervals based on the breastfeeding example where interest is in estimating the time until weaning. These intervals were calculated for the adjusted NPMLE using classification rates of $\alpha = 0.9$ and $\beta = 1$. With a sample size of $n = 8041$, the corresponding block sizes range from 20 to 3271. Table A.2 shows similar results for the menopause example where interest is in estimating the time until natural menopause, under the classification rates of $\alpha = 0.95$ and $\beta = 0.9$. Given a sample size of $n = 2076$, the block sizes range from 13 to 967.

t_0	46.5 years	49.5 years	51.5 years
$\hat{F}(t_0)$	0.300	0.578	0.738
$m = 13$ $(n^{1/3})$	[0.167 0.433]	[0.468 0.688]	[0.597 0.879]
$m = 46$ $(n^{1/2})$	[0.214 0.386]	[0.413 0.743]	[0.558 0.918]
$m = 163$ $(n^{2/3})$	[0.167 0.433]	[0.351 0.805]	[0.558 0.918]
$m = 307$ $(n^{3/4})$	[0.135 0.465]	[0.343 0.813]	[0.581 0.895]
$m = 450$ $(n^{0.8})$	[0.112 0.488]	[0.319 0.837]	[0.573 0.903]
$m = 967$ $(n^{0.9})$	[0.073 0.527]	[0.304 0.852]	[0.581 0.895]

Table A.2: Menopause example: estimation of the time until natural menopause. Confidence interval estimation for the adjusted ($\alpha = 0.95$, $\beta = 0.9$) NPMLE at three monitoring times obtained using the m out of n bootstrap for various values of m ranging from 13 to 967.

For Table A.1 and Table A.2, the results are given for three monitoring times, with the corresponding estimate of F at these monitoring times also given. These estimates of F are based on the values of the NPMLE adjusted for constant misclassification. Both tables suggest that the 95% symmetric confidence intervals are quite stable across the presented monitoring times, which were chosen to represent the overall spread of monitoring times. For each choice of block size m, the results are based on 500 bootstrap samples. Similar results were also found for a larger number of bootstrap samples. Overall, these tables, and the corresponding table of Section 2.2.2 (Table 2.1) suggest the m out of n bootstrap can be used to provide a clear, consistent and useful assessment of variability.

A.2.2 Time-varying Misclassification

We now extend this simple misclassification scenario, where misclassification is constant and independent of time, to allow for the possibility of classification rates that vary with time. Recall the specific time-varying misclassification models of Section 2.2.3 where misclassification is highest when the observed monitoring time is close to the true failure event. Under such a time-varying misclassification model, there is currently no available method for direct estimation of F. However, although they do not represent the time-varying misclassification model exactly, there are two available alternative methods, namely, the unconstrained NPMLE and the NPMLE adjusted for constant response misclassification, which can be used to estimate F. In the following tables we therefore compare these two estimators to determine which approach would be most accurate if the data is truly misclassified according to this particular time-varying misclassification model.

In Section 2.2.3 we examined two related time varying misclassification models, both following the same general structure. First, misclassification was assumed to occur only within a known time interval surrounding the true failure event, with perfect sensitivity and specificity observed for all monitoring times outside this interval. Table 2.3 gives estimates for both the unconstrained NPMLE and the NPMLE adjusted for constant misclassification, under this scenario. An extension of this time varying misclassification model was then considered, where all individuals are subject to misclassification, but the classification rates are dependent on how close the monitoring time (C) is to the true failure time (T). For a defined length A, higher rates of misclassification are applied if $|C - T| \leq A$. Otherwise, the current status outcomes are still subject to misclassification, but at lower rates of misclassification. The results of this extension are given in Table 2.4 where estimates of both the unconstrained NPMLE and adjusted NPMLE are given for various values of the interval length, A. In this appendix we re-examine both of these misclassification models where the menopause example is most appropriately described by the first model, with the breastfeeding example following the extension. We therefore have the ability to present additional results similar to those of both Table 2.3 and Table 2.4. Table A.3 and Table A.4 show the results of both estimators of F for the breastfeeding and menopause examples, respectively.

C		6 months	15 months	20 months	29 months	39 months
		$F(C)$ =0.123	$F(C)$ = 0.263	$F(C)$ = 0.457	$F(C)$ = 0.688	$F(C)$ = 0.833
$A = 0$	0% (0)%					
	NPMLE$_0$	0.114(0.018)	0.262(0.020)	0.445(0.037)	0.689(0.017)	0.844(0.011)
	NPMLE$_\infty$	0.126(0.020)	0.291(0.022)	0.494(0.042)	0.766(0.019)	0.937(0.012)
Bias	NPMLE$_0$	0.120(0.025)	0.271(0.026)	0.473(0.047)	0.713(0.027)	0.867(0.018)
adjusted	NPMLE$_\infty$	0.120(0.025)	0.271(0.026)	0.473(0.048)	0.711(0.028)	0.865(0.018)
$A = 5$	17% (3.7)%					
	NPMLE$_0$	0.112(0.019)	0.254(0.021)	0.441(0.040)	0.685(0.024)	0.835(0.013)
	NPMLE$_\infty$	0.125(0.021)	0.282(0.024)	0.490(0.045)	0.761(0.026)	0.928(0.015)
Bias	NPMLE$_0$	0.119(0.026)	0.278(0.027)	0.473(0.052)	0.713(0.026)	0.869(0.019)
adjusted	NPMLE$_\infty$	0.119(0.026)	0.278(0.027)	0.474(0.053)	0.712(0.027)	0.867(0.020)
$A = 15$	46% (4.5)%					
	NPMLE$_0$	0.109(0.017)	0.251(0.022)	0.432(0.037)	0.665(0.019)	0.826(0.013)
	NPMLE$_\infty$	0.121(0.018)	0.279(0.025)	0.480(0.041)	0.739(0.021)	0.917(0.015)
Bias	NPMLE$_0$	0.115(0.022)	0.270(0.025)	0.461(0.053)	0.703(0.027)	0.870(0.018)
adjusted	NPMLE$_\infty$	0.116(0.023)	0.271(0.025)	0.463(0.054)	0.704(0.027)	0.869(0.019)
$A = \infty$	100% (6.4)%					
	NPMLE$_0$	0.106(0.017)	0.251(0.018)	0.423(0.038)	0.656(0.023)	0.796(0.014)
	NPMLE$_\infty$	0.118(0.019)	0.279(0.020)	0.470(0.042)	0.729(0.025)	0.884(0.015)
Bias	NPMLE$_0$	0.119(0.021)	0.264(0.026)	0.462(0.042)	0.693(0.026)	0.846(0.018)
adjusted	NPMLE$_\infty$	0.119(0.022)	0.264(0.027)	0.463(0.043)	0.693(0.027)	0.846(0.019)

Table A.3: Breastfeeding example: estimation of time to weaning. Simulation averages (standard deviations) of two estimators of F at 5 monitoring times, when the data generating distribution is subject to classification rates of $\alpha = 0.9$, $\beta = 1$ within a window of length $2A$ surrounding the failure time, and $\alpha = 0.95$, $\beta = 1$ elsewhere. The resulting % subject to misclassification (average % actually misclassified) are also given. NPMLE$_0$ and NPMLE$_\infty$ represent the unconstrained NPMLE and the NPMLE adjusted for constant misclassification, respectively.

Table A.3 shows the estimates for the unconstrained NPMLE (NPMLE$_0$) and the adjusted NPMLE (NPMLE$_\infty$) for the breastfeeding example, where interest is in estimating the time until weaning. The data is truly misclassified with the classification rates of $\alpha = 0.9$ and $\beta = 1$ for women examined within a pre-defined length A from the true failure event T, and at classification rates of $\alpha = 0.95$ and $\beta = 1$ for those examined more than A units from T. A selection of monitoring times from 6 months to 39 months are presented. A variety of values of A are considered which give intervals of different lengths in which the higher rates of misclassification are observed. Values of A representing no misclassification ($A = 0$) and constant misclassification ($A = \infty$) are also presented. For the NPMLE adjusted for constant misclassification (NPMLE$_\infty$), constant classification rates of $\alpha = 0.9$ and $\beta = 1$ are assumed. The results of this table can be compared with those of Table 2.4.

Table A.4 gives similar results for both estimators of F, for the menopause example where interest is now on estimating the time until natural menopause. In this example, the data is only subject to misclassification if the monitoring time occurs in the interval $[T - A, T + A]$, for an appropriate choice of A. Perfect classification is assumed for all monitoring times outside this window. In this table the NPMLE adjusted for constant misclassification assumes constant rates of classification of $\alpha = 0.95$ and $\beta = 0.9$, which are the true classification rates within the interval $[T - A, T + A]$. A selection of monitoring times and values of A defining the length of the interval are again evaluated and presented in the table, including $A = 0$ (no misclassification) and $A = \infty$ (constant misclassification). The results of this table can be compared with those of Table 2.3. The total percentage of individuals subject to misclassification (average percentage actually misclassified) are also given in Table A.3 and Table A.4, for each assumed value of A.

In both Table A.3 and Table A.4 it can be seen that due to the high levels of classification, even when all individuals are subject to misclassification, only an average of 6.4% and 9.0% of individuals are actually misclassified, respectively. When all individuals are subject to misclassification ($A = \infty$), we would therefore expect the unconstrained NPMLE of Tables A.3 and A.4 to perform better than the corresponding unconstrained NPMLE of Table 2.3, where an average of 20% of individuals would actually be misclassified. In Table A.3 the greatest difference between the two estimators (NPMLE$_0$ and NPMLE$_\infty$) is seen at the later monitoring times. This result is not surprising, given Figure A.1(a). From this figure it is seen that since $\beta = 1$, the difference between the curves is driven by the imperfect classification $\alpha = 0.9$ which has a greater impact at the later monitoring times, when the estimate of F is closer to 1. The reverse is seen in Table A.4 where the greatest difference between the two estimators is seen at the earlier monitoring times. This is a result of the sensitivity being greater than the specificity. As was seen in Chapter 2, both estimators produce biased estimates of F. Again, we apply the bias-adjusted approach described in Section 2.2.3, the results of which are given in Tables A.3 and A.4 for each of the assumed values of A, including $A = 0$ and $A = \infty$. In both tables, the bias-adjusted estimates are extremely similar for both estimators.

C		45.5 years	48.5 years	50.5 years	52.5 years	54.5 years
		$F(C)$	$F(C)$	$F(C)$	$F(C)$	$F(C)$
		$=0.152$	$= 0.295$	$= 0.593$	$= 0.789$	$= 0.875$
$A = 0$	0% (0)%					
	NPMLE$_0$	0.144(0.037)	0.307(0.049)	0.583(0.063)	0.771(0.052)	0.875(0.048)
	NPMLE$_\infty$	0.054(0.041)	0.243(0.057)	0.569(0.074)	0.789(0.061)	0.911(0.054)
Bias	NPMLE$_0$	0.152(0.045)	0.297(0.058)	0.600(0.077)	0.779(0.061)	0.881(0.059)
adjusted	NPMLE$_\infty$	0.139(0.061)	0.302(0.059)	0.604(0.078)	0.779(0.063)	0.882(0.065)
$A = 4.5$	22% (1.6)%					
	NPMLE$_0$	0.165(0.035)	0.341(0.050)	0.594(0.061)	0.774(0.048)	0.871(0.046)
	NPMLE$_\infty$	0.077(0.041)	0.283(0.059)	0.581(0.072)	0.793(0.056)	0.906(0.052)
Bias	NPMLE$_0$	0.147(0.045)	0.300(0.065)	0.591(0.079)	0.771(0.074)	0.871(0.067)
adjusted	NPMLE$_\infty$	0.149(0.049)	0.298(0.067)	0.592(0.081)	0.770(0.077)	0.870(0.071)
$A = 10$	48% (3.9)%					
	NPMLE$_0$	0.144(0.037)	0.307(0.049)	0.583(0.063)	0.771(0.052)	0.875(0.048)
	NPMLE$_\infty$	0.054(0.041)	0.243(0.057)	0.569(0.074)	0.789(0.061)	0.911(0.054)
Bias	NPMLE$_0$	0.152(0.054)	0.303(0.063)	0.581(0.076)	0.769(0.070)	0.866(0.067)
adjusted	NPMLE$_\infty$	0.147(0.056)	0.297(0.065)	0.579(0.079)	0.769(0.073)	0.866(0.070)
$A = \infty$	100% (9.0)%					
	NPMLE$_0$	0.183(0.055)	0.334(0.055)	0.590(0.062)	0.766(0.051)	0.863(0.048)
	NPMLE$_\infty$	0.099(0.062)	0.276(0.065)	0.577(0.072)	0.783(0.061)	0.897(0.055)
Bias	NPMLE$_0$	0.150(0.056)	0.303(0.060)	0.581(0.084)	0.751(0.079)	0.859(0.071)
adjusted	NPMLE$_\infty$	0.144(0.058)	0.296(0.062)	0.579(0.087)	0.749(0.082)	0.857(0.074)

Table A.4: Menopause example: estimation of time to natural menopause. Simulation averages (standard deviations) of two estimators of F at 5 monitoring times, when the data generating distribution is subject to classification rates of $\alpha = 0.95$, $\beta = 0.9$ within a window of length $2A$ surrounding the failure time, and $\alpha = 1$, $\beta = 1$ elsewhere. The resulting % subject to misclassification (average % actually misclassified) are also given. NPMLE$_0$ and NPMLE$_\infty$ represent the unconstrained NPMLE and the NPMLE adjusted for constant misclassification, respectively.

A.2.3 Regression

The correspondence between standard regression models for the underlying failure time and generalized linear models for the observed current status outcome was described in Section 2.2.5, along with a description of how to allow for misclassification through modification of the link function. For convenience we repeat the modified link function here

$$g^*(P(\Delta = 1|X, C)) = g\left\{\frac{P(\Delta = 1|X, C) - (1 - \beta)}{(\alpha + \beta - 1)}\right\}.$$

Keeping consistent with Chapter 2, we assume a Weibull regression model for T, which results in the generalized linear model for Y in the covariate X and the monitoring time C that involves g, the complementary log-log link function. Although the National Health Care Surveys contain a wide variety of information, unfortunately the menopausal data available to us does not contain any covariate information. We therefore only consider the breastfeeding example as an additional application in the regression setting.

In the breastfeeding application we are interested in estimating the time until weaning. We now consider the mother's residence as a covariate of interest, where residence is recorded as a binary variable indicating residence in a rural (residence = 0) or urban (residence = 1) location. We fit regression models to this breastfeeding data (a) assuming no errors in the response variable (therefore using g directly), and (b) adjusting for errors with constant classification rates of $\alpha = 0.9$ and $\beta = 1$ (using g^*). The parameter estimates in both models have a proportional hazards interpretation on age at weaning, according to the Weibull regression model assumption for T. The age of the child at screening is again included in the model additively on the log scale. Table A.5 presents the results of models (a) and (b), along with the observed ratio of parameter estimates.

Covariate	log (RH): Model (a) ($\hat{\beta}^*$) Ignoring misclassification	log(RH): Model (b) ($\hat{\beta}$) Adjusted for misclassification	$\hat{\beta}^*/\hat{\beta}$
Residence	0.238 (0.036)	0.340 (0.036)	0.7
log(age at screening)	1.512 (0.034)	1.366 (0.011)	1.1

Table A.5: Estimates (and standard errors) of the log Relative Hazard (RH) for time to weaning, which is assumed to follow a Weibull distribution. Model (a) ignores misclassification in the response variable and Model (b) incorporates constant misclassification corresponding to $\alpha = 0.9$ and $\beta = 1$.

According to models (a) and (b), the hazard of weaning is significantly increased by 26% and 40% for those who have an urban residence (Residence = 1) to those who have a rural residence (Residence = 0), respectively, holding other covariates in the model fixed. The ratio of the parameter estimates suggest ignoring errors in the response variable leads to substantially biased estimates of the association of covariates with breastfeeding status.

Appendix B

Human Papillomavirus (HPV)

This appendix aims to give the reader a general understanding of various aspects of the Human Papillomavirus (HPV), including information on HPV prevalence, biological features, prevention and detection. This overview is by no means complete, its goal is to help motivate and give a deeper understanding of the importance of the methods outline in Chapter 2. Knowledge of the infection of interest is not necessary for understanding the statistical concepts presented in this dissertation, but gives a more complete assessment of the work. Unless stated otherwise, the information on HPV contained within this appendix is based on the best available information known to the Centers for Disease Control and Prevention (CDC) in 2011 (CDC, 2011b).

B.1 Introduction and Prevalence

Genital Human Papillomavirus, most commonly referred to as simply HPV, is a very common, usually benign, sexually transmitted infection. HPV affects skin and mucous membranes, and is the cause of warts. Approximately 100 viral types of HPV have been identified, and about one third of these are associated with sexually transmitted genital infections (Koutsky and Kiviat, 1999). These HPV types can also infect the mouth and throat. HPV is not the same as HIV (see Appendix C), both are viruses that can be passed on during sex but they cause different symptoms and health problems.

HPV has affected humans for thousands of years and is transmitted by direct skin to skin contact with an infected individual. Transmission is usually from vaginal, oral or anal sexual contact, and can occur whether or not warts or other symptoms are present (McDermott-Webster, 1999). HPV can be passed on between straight and same-sex partners. A person can have HPV even if years have passed since he or she had sexual contact with an infected person. It is possible to get more than one type of HPV. The virus, although it rarely occurs, can also be transmitted from mother to infant during childbirth. This vertical transmission is associated with the child's development of recurrent laryngeal papillomatosis (warts on the throat). This type of transmission occurs in approximately 2,000 children for every 4 million new borns (Jay and Moscicki, 2000). This is a serious condition that may require frequent laser surgery to prevent obstruction of the infant's airways.

HPV is the most common sexually transmitted infection in the United States. According to the American Social Health Association (ASHA, 2011), approximately 5.5 million new cases of sexually transmitted HPV infections are reported every year, although 6.2 million new infections are estimated to occur. As most people with HPV do not develop symptoms or health problems from it, although approximately 20 million people in the United States are currently infected with HPV, most of them are unaware they are infected, and many have never even heard of the virus. HPV is so common that at least 50% of sexually active men and women get it at some point in their lives. The highest rates of genital HPV infection are found in adults between the ages of 18 and 28 (Koutsky, 1997). HPV is believed to be widespread across racial groups with very little variation in prevalence across regions (CDC, 2000). There is no treatment for the virus itself, but there are treatments for the diseases that HPV can cause.

As described in Section B.5, genital warts or cancer can develop from HPV infection. About 1% of sexually active adults in the U.S. have genital warts at any one time. Each year, about 12,000 women get cervical cancer in the U.S. Almost all of these cancers are HPV-associated. For less common cancers, each year in the United States, approximately 1,500 women get HPV-associated vulvar cancer, 500 women get HPV-associated vaginal cancer, 400 men get HPV-associated penile cancer, 2,700 women and 1,500 men get HPV-associated anal cancer, 1,500 women and 5,600 men get HPV-associated oropharyngeal cancers (cancers of the back of throat including base of tongue and tonsils). Although each of these cancers occur less frequently than cervical cancer, taken together they equal nearly half the number of cases of cervical cancer in the U.S. (Eng and Butler, 1997). It must be noted that certain populations are at higher risk for some HPV-related health problems. These include gay and bisexual men, and people with weak immune systems (including those who have HIV/AIDS).

B.2 Biological Features

HPV infection can be clinical (symptomatic) or subclinical (asymptomatic), and many people with HPV will never know they have it. HPV targets the deep, basal level of the skin and most often causes no clinical microscopic changes in the cells of the skin (Verdon, 1997). Knowledge of the HPV virus has progressed over the years. A recent paper by Schiller et al. (2010) gives a current understanding of the mechanism of HPV infection, where it is stated that HPV has the unique mechanism of infection that has likely evolved to limit infection to the basal cells of stratified epithelium, the only tissue in which they can replicate. The virus cannot bind to live tissue, instead, it infects epithelial tissues through microabrasions or other epithelial trauma that exposes segments of the basement membrane. The infectious process is slow, taking 1224 hours for initiation of transcription. It is believed that involved antibodies play a major neutralizing role while the virions still reside on the basement membrane and cell surfaces.

B.3 Detection

HPVs cannot be cultured and the detection of HPV relies on a variety of techniques used in immunology, serology, and molecular biology (Villa and Denny, 2006). There is currently no FDA-approved test to detect HPV in men but HPV tests are available for women and are strongly recommended at regular intervals. Current HPV tests can detect up to 13 high-risk and 5 low-risk HPV types (Hybrid Capture assay, version h2). Villa and Denny (2006) discuss the advantages and disadvantages of the different methods of HPV DNA detection.

Most people with HPV do not develop symptoms or health problems from it. In 90% of cases, the body's immune system clears HPV naturally within two years (CDC, 2011b). In a study designed to determine the natural history of genital HPV infection, college women were followed for three years and HPV was detected using a sensitive DNA test designed to detects small amounts of HPV, even when there are no symptoms present. It was reported that 43% tested positive for HPV at some point over the study period. Repeated HPV DNA testing showed that 70% of the HPV infected women cleared their HPV infections within one year through the natural immune process, with only 9% still infected after two years. The viral type of HPV was found to be a major determinant in the duration of infection, with types 16, AE7, 61, 18, and 73 having the longest average duration (Ho et al., 1998).

B.4 Vaccines and Prevention

In 90% of cases, the body's immune system clears HPV infection naturally within two years without any medical intervention. If HPV infections are not cleared they can develop into genital warts or cancer, including cervical cancer and other less common but serious cancers. There is no way to know which people who get HPV will go on to develop cancer or other health problems (CDC, 2011b). Although there is no cure for HPV, there are several ways in which people can lower their chance of getting HPV. Vaccines exist that can protect males and females against some of the most common types of HPV that can lead to disease and cancer. These vaccines are given in three shots and are most effective when given at

11 or 12 years of age. There are two vaccines (Cervarix and Gardasil) available to protect females against the types of HPV that cause most cervical cancers but there is only one available vaccine (Gardasil) to protect males against most genital warts and anal cancers. HPV vaccines will not treat or get rid of existing HPV infections. Also, HPV vaccines do not treat or cure health problems (like cancer or warts) caused by an HPV infection that occurred before vaccination. All vaccines used in the United States are required to go through years of extensive safety and testing before they are licensed by the U.S. Food and Drug Administration (FDA). Both HPV vaccines, Cervarix and Gardasil, are currently being monitored for adverse events, especially rare events not identified in the study trials.

Condoms may lower the risk of HPV or HPV-related diseases, such as genital warts and cervical cancer, but HPV can also infect areas that are not covered by a condom, so condoms are not fully protective against HPV. People can also lower their chance of getting HPV by being in a faithful relationship with one partner, limiting their number of sexual partners, and choosing a partner who has had no or few prior sex partners. However, even people with only one lifetime sexual partner can get HPV, and it may not be possible to determine if a partner who has been sexually active in the past is currently infected.

Cervical cancer can also be prevented with routine cervical cancer screening (pap test) and follow-up of abnormal results. The pap test can find abnormal cells on the cervix so that they can be removed before cancer develops. Abnormal cells often become normal over time, but can sometimes turn into cancer. These cells can usually be treated, depending on their severity, the woman's age, past medical history, and other test results. HPV DNA tests may also be used with a pap test in certain cases. Even women who were vaccinated when they were younger need regular cervical cancer screening because the vaccines do not protect against all cervical cancers.

B.5 Warts and Cancer

HPV can cause normal cells on the infected skin to turn abnormal. Most of the time, you cannot see or feel these cell changes. In most cases, the body fights off HPV naturally and the infected cells then go back to normal. In cases when the body does not fight off HPV, HPV can cause visible changes in the form of genital warts or cancer. Warts can appear within weeks or months after sexual contact with an infected partner, even if the infected partner has no signs of genital warts. Cancer, however, often takes years to develop.

The types of HPV that can cause genital warts are not the same as the types that can cause cancers. Genital warts usually appear as a single bump or group of bumps in the genital area. They can be small or large, raised or flat, or shaped like a cauliflower. Warts can be diagnosed by looking at the genital area during a medical examination. If left untreated, genital warts might go away, remain unchanged, or increase in size or number, but they will not turn into cancer. Other types of HPV can cause common warts in other areas such as the hands or feet. Recurrent Respiratory Papillomatosis (RRP) causes warts to grow in the throat. This can sometimes block the airway, causing a hoarse voice or troubled breathing. Although rare, RRP can occur among adults or children.

Genital warts are the clinical, visible manifestation of genital HPV. In more than 90% of cases, they are caused by HPV types 6 and 11, which are considered low-risk types because they are not associated with an increased risk of cancer. However, a person may be infected with more than one type of HPV at a single point in time. It is estimated that 1% of the American population has genital warts, with men and women having similar rates. Genital warts are very contagious, with an estimated rate of infection between 30% and 60% from unprotected exposure (Jay and Moscicki, 2000).

Cervical cancer usually does not have symptoms until it is quite advanced. For this reason, it is important for women to get regular screening for cervical cancer. Screening tests can find early signs of the disease so that problems can be treated early, before they ever turn into cancer. An international study found that HPV was present in 93% of cervical cancer tumors. HPV 16 was found in 50% of these cases, HPV 18 accounted for 14%, HPV 45 for 8%, HPV 31 for 5% and other types accounted for 23% of all cases (Bosch et al., 1995). Although HPV is considered a cause of cervical cancer, only 1 out of 1000 women with HPV develop invasive cervical cancer. HPV appears to be necessary, but not sufficient, for the development of cervical cancer. Apart from HPV type, researchers believe there are several other factors that may contribute to the development of cervical cancer. These factors may include smoking, HIV infection, diet, hormonal factors, and the presence of other sexually transmitted infections, such as chlamydia or herpes simplex virus 2 (Anttila et al., 2001). Other HPV-related cancers might not have signs or symptoms until they are advanced, at which time they will be difficult to treat. These include cancers of the vulva, vagina, penis, anus, and oropharynx (back of throat including base of tongue and tonsils).

Appendix C

Human Immunodeficiency Virus (HIV)

This appendix is similar to Appendix B (Human Papillomavirus) and seeks to give the reader a general understanding of various aspects of the Human Immunodeficiency Virus (HIV), including information on HIV prevalence, biological features, prevention and detection. Again, this overview is not complete, but gives enough information to help convey the importance of the methods outlined in Chapter 4. The HIV testing mechanism used in the illustration of Chapter 4 is given in more detail in Section C.3. Unless stated otherwise, the information contained within this appendix, relating to Human Immunodeficiency Virus and Acquired Immune Deficiency Syndrome (AIDS), is based on the best available information known to the Centers for Disease Control and Prevention (CDC) in 2011 (CDC, 2011a), where several studies published in scientific journals are listed, from which many of the details presented here were obtained.

C.1 Introduction and Prevalence

Although there are two types of HIV (HIV-1, HIV-2), in much of the literature, and also in this appendix, HIV-1 is simply referred to as HIV. The HIV virus infects the white blood cells, known as CD4+ cells, which are part of the body's immune system that help fight infections. The CDC estimates that about 56,300 people in the United States contracted HIV in 2006 (Hall et al., 2008). Despite the major advances in diagnosing and treating HIV infection, 35,962 cases of AIDS were diagnosed in 2007, and 14,110 deaths were reported among people living with HIV in the United States. Untreated early HIV infection is also associated with many diseases including cardiovascular disease, kidney disease, liver disease, and cancer. It is currently estimated that approximately 1.1 million Americans are now living with HIV (NIAID, 2011). Once infected with HIV, the virus can later progress to AIDS, which is a serious syndrome, more information on which is given in Section C.5.

Although there are many less common modes of HIV transmission, including blood transfusions and unsanitary injections, the primary forms of HIV transmission come from unprotected sex with a person who has HIV, sharing needles, and being born to an infected mother, who can spread HIV during pregnancy, birth or breastfeeding. Unprotected sex with an HIV infected person has various levels of risk. The highest risk is associated with having unprotected anal sex, multiple sex partners or the presence of other sexually transmitted diseases. Although this feature of HIV is commonly misunderstood, HIV cannot reproduce outside the human body, therefore HIV cannot be spread by air or water, by insects (including mosquitos), saliva, sweat, tears, or casual contact such as shaking hands.

C.2 Biological Features

According to the National Institute of Allergies and Infectious Diseases in 2011 (NIAID, 2011), HIV belongs to a class of viruses known as retroviruses, and more specifically to a subgroup known as lentiviruses. This type of virus has a long waiting time period between initial infection and the beginning of serious symptoms. Retroviruses are viruses that contain RNA (ribonucleic acid) as their genetic material. After infecting a cell, HIV uses an enzyme called reverse transcriptase to convert its RNA into DNA (deoxyribonucleic acid), and then proceeds to replicate itself using the cell's machinery. A very useful and detailed description of the biology of HIV, including the structure of HIV and the HIV replication cycle can be found on the National Institute of Allergies and Infectious Diseases' website (NIAID, 2011).

HIV replicates rapidly with several billion new viruses made every day in a person infected with HIV. Therefore, HIV is difficult to stop due to its ability to mutate and evolve rapidly. New strains of HIV develop in HIV infected individuals, some strains of which are easy to kill while others can replicate at faster rates, making them more difficult to kill. The more virulent and infectious strains of HIV are typically found in people who are in the late stages of infection. The development of AIDS also occurs in the late stage of HIV infection (Section C.5). Different strains of HIV can also recombine to produce an even wider range of strains. One of the effects of HIV is that the virus destroys billions of CD4+ cells in a person infected with HIV (per day), which eventually prevents the immune system from regenerating or fighting other infections. HIV can be shielded from the immune system by lying dormant in an infected cell for months or even years. Antiretroviral drugs are capable of suppressing HIV, even to undetectable levels in the blood, but they cannot eliminate the virus hiding in these cells.

Recall Figure 4.5 of Chapter 4, reproduced below (Figure C.1) for convenience. This figure gives a basic schematic to show the progression of HIV over time, where the time scale begins at the time of initial HIV infection. Once HIV enters the body, the virus infects a large number of CD4+ cells and replicates rapidly. During this acute phase of infection, the blood has a high number of HIV copies (viral load) that spread throughout the body, which can be seen in Figure C.1. After infection, it can take weeks to months for HIV specific antibodies to appear in the blood sample. The time between being infected with HIV and the time at which these HIV specific antibodies can be detected is called the seroconversion, during which, an infected individual can still spread the disease, even though a test will not indicate an HIV positive result.

As was utilized in Chapter 4, the detection of certain biomarkers can be used to indicate a particular disease state. We can see from Figure C.1 that the level of antibodies present for someone infected for a few days differs greatly from the corresponding levels for someone infected for several years. As described in Section 4.4, Janssen et al. (1998) make use of this feature of HIV progression to identify those recently infected, knowledge of which is important for clinical care and prevention. Note also that in Chapter 4 we are interested in estimating the time until an HIV positive result on a sensitive diagnostic test and not in the time until HIV infection. Figure C.1 helps illustrate how an antibody test, such as the ELISA, described below in Section C.3, may not detect HIV specific antibodies in someone who has been recently infected with HIV as there are not yet enough detectable antibodies in the blood sample. This again supports the time-varying misclassification approach considered in Section 2.2.3 of Chapter 2.

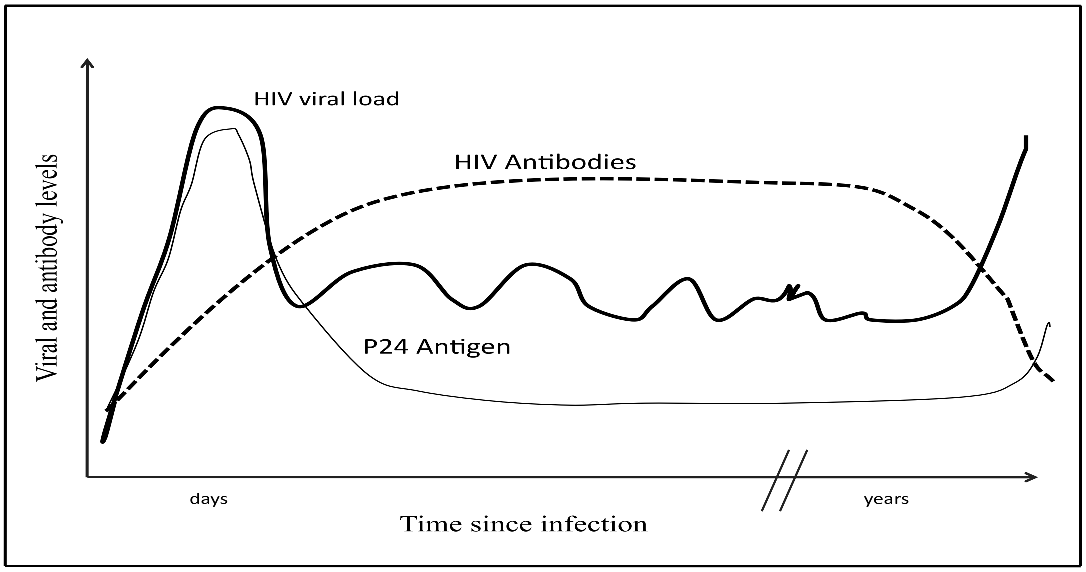

Figure C.1: Reproduction of Figure 4.5: Schematic of HIV progression over time.

C.3 Detection

In the acute phase of infection, up to 70% of HIV infected individuals suffer flu-like symptoms (NIAID, 2011). Other people infected with HIV may not show any symptoms at all. However, although these infected individuals may both feel and appear healthy for several years, HIV is still damaging their body. Each year, approximately 16-22 million persons in the United States are tested for HIV. By 2002, an estimated 38%-44% of all adults had been tested for HIV (CDC, 1999). However, at the end of 2006, approximately 1 in 5 (21%, or 232,700 persons) persons did not know they were infected (CDC, 2008). An individual's knowledge of their HIV status is important for maintaining health and reducing the spread of the virus. In 2006, in an effort to increase such knowledge, the CDC implemented new recommendations for HIV testing for both inpatients and outpatients in acute-care hospital settings. In this recommendation, the routine HIV screening of adults, adolescents, and pregnant women in health care settings in the United States was advised. Once aware of being HIV infected, people generally take advantage of the therapies aimed to keep them healthy and extend their lives. Without knowledge of being infected with HIV, people are unlikely to alter their behaviors that allow transmission of HIV. Several cohort studies have shown that many infected persons decrease behaviors that transmit infection to sexual or needle-sharing partners once they are aware of their positive HIV status (Rietmeijer et al., 1996, Rhodes and Malotte, 1996, Gibson et al., 1999). Medical treatment designed to lower HIV viral load may also reduce the risk of transmission to others (Quinn et al., 2000). Therefore, early referral to medical care could prevent HIV transmission in communities while also reducing a person's risk for HIV-related illnesses and death. A sample of blood can be tested for the presence of HIV specific antibodies (disease-fighting proteins). Groups or batches of blood samples can also be tested for the presence of at least one infected individual sample, as described in Chapter 3.

Enzyme-Linked Immunosorbent Assay (ELISA)

The Enzyme-Linked Immunosorbent Assay (ELISA), introduced in Section 4.4, is a technique used to detect the presence of an antibody or antigen in a sample. It can be used for a variety of applications, and is a useful tool for determining HIV antibody concentrations. The ELISA is usually the first test performed to detect infection with HIV. The test is generally repeated to confirm diagnosis, only when an initial positive result is obtained. The ELISA has a low chance of producing a false result after the first few weeks of infection, that is, after seroconversion. Due to its high sensitivity level, the ELISA was the first screening test widely used for HIV detection. ELISA results are reported as a number. The most controversial aspect of this test is determining the cut-off point between a positive and a negative result. Once the cut-off is determined, samples that generate a signal stronger than the known sample (prepared at the cut off rate) are considered positive and samples that generate weaker signals are considered negative. Lequin (2005) describes the historical background of the invention of the ELISA.

C.4 Vaccines and Prevention

Recent developments have greatly aided the prevention of HIV transmission and the overall care available to individuals infected with HIV. The Food and Drug Administration have approved several medications (31 antiretroviral drugs) to treat HIV infection which can limit or slow down the destruction of the immune system. These medications can also improve the health of people living with HIV, and may reduce their ability to transmit the virus (CDC, 2011a). While current medications can dramatically improve the health of people living with HIV and slow down the progression of HIV infection to AIDS, existing treatments need to be taken daily for the rest of a person's life, need to be carefully monitored, and come with costs and potential side effects. Available treatments can suppress the virus, even to undetectable levels, but cannot completely eliminate HIV from the body (NIAID, 2011). At this time, there is no cure for HIV infection.

Many agencies, including the National Institute of Allergies and Infectious Diseases, conduct and support research to better understand HIV and how it causes disease, with the hope of finding new and more effective therapies, drug classes, and antiretroviral drug combinations that can extend and improve the quality of life for people living with HIV and AIDS (NIAID, 2011). Although there is no vaccine to prevent the development of HIV, there are many lifestyle changes and preventative measures one can practice to reduce the risk of becoming infected with HIV, or transmitting the virus once infected. Such measures include being regularly tested for HIV, remaining faithful to your partner, using condoms during sex, and not sharing needles.

C.5 Acquired Immune Deficiency Syndrome (AIDS)

Acquired Immune Deficiency Syndrome (AIDS) is a disease of the human immune system caused by the HIV virus (Sepkowitz, 2001). As HIV attacks the CD4+ cells, a person's immune system is weakened and they are unable to fight off infection and disease, ultimately resulting in the development of AIDS. AIDS is therefore the late stage of HIV infection, when a person's immune system is severely damaged and has difficulty fighting diseases and certain cancers. Significant developments in available medications, since AIDS was first recognized in 1981, have allowed people to live much longer before they develop AIDS, a process that once took only a few years (CDC, 2011a). AIDS remains actively spreading and is considered a severe health problem in many parts of the world. An epidemic update published by the Joint United Nations programme on HIV and AIDS, reported that, in 2007, an estimated 33.2 million people worldwide were living with AIDS, and that 2.1 million people died from AIDS during that year, including 330,000 children. It is estimated that 76% of those deaths occurred in sub-Saharan Africa (WHO, 2007).