Statistical Methods for Interval Censored Data with Informative Sampling Weights and Unknown Subpopulation Not at Risk for the Event *

Aiko Hattori, Ph.D.

University of California, Los Angeles

Chirayath M. Suchindran, Ph.D.

University of North Carolina at Chapel Hill

* Please note that this is a preliminary report of the study. Further work is planned (as discussed in the discussion section) and the results will be included in the final version of this paper to be uploaded on the PAA 2012 program website by April 2, 2012.

SUMMARY

Research on interval censored time to event using complex survey data faces three methodological challenges: clustered data with informative sampling weights, interval-censored event times, and an unknown subpopulation not at risk. This paper has two study aims. First we use non-parametric maximum likelihood methods using Turnbull algorithm to estimate the Kaplan-Meier analog of survival function with this data. We further extend the Pseudo Maximum Likelihood method to a random effect, mixture distribution model within the framework of Accelerated Failure Time model. Particular attention is paid to the scaling of sample weights in multistage sampling. We use the National Longitudinal Study of Adolescent Health data and assess time to obesity among adolescents. We confirm that the magnitude of regression coefficients can be biased when informative sampling design is not accounted for. Also the variances are underestimated when clustering is ignored. When the two different subpopulations, one at risk and the other not at risk for the event, are not addressed in a model, the regression coefficients may be biased and mask the true association between covariates and time to event.

1. INTRODUCTION

Data collected through sample surveys have been widely used for hypothesis testing through different statistical methods including survival analysis. Survival analysis, a statistical method to examine time to event of interest and its association with risk factors, may face methodological challenges when analyzing complex survey data. Major methodological challenges lie in that: 1) data collected through multistage sampling survey are often clustered and have unequal sampling probabilities associated with time to event of interest; 2) the exact time to event may not be observed by researchers; and 3) there may be an unknown subpopulation not at risk for the event of interest.

Multistage sampling designs often cause clustering, i.e., correlation of observations at a lower stage of sampling within higher sampling units (Kish 1965; Lohr 1999). When clustering is ignored, variance estimates of parameters are often biased downward. Another problem that data collected through multistage sampling designs may possess is unequal sampling probabilities associated with time to event of interest. When sampling probabilities at one or more sampling stages depend on time to event of interest after accounting for covariates and vary across a sample, the sampling design is said to be informative. Informative sampling design causes standard estimators of parameters to be biased (Grilli and Pratesi 2004; Korn and Graubard 2003; Pfeffermann et al. 1998; Skinner 1989). The direction and magnitude of the bias cannot be determined as a priori knowledge.

To address the bias when sampling design is informative, Skinner (1989) presents several alternative design-based methods, including the Pseudo Maximum Likelihood (PML) and Moment Structures and Generalized Least-Squares (GLS) methods. Pfeffermann *et al*. (1998) propose a statistical approach to incorporate sampling design in the context of a two-stage sampling design, following the work by Skinner (1989). They prove that when the sample size is sufficiently large the PML method applying reciprocals of the sampling probabilities at each sampling stage can yield consistent estimators of parameters in a model with a continuous dependent variable. Grilli and Pratesi (2004) extend the work by Pfeffermann *et al*. (1998) and Skinner (1989), and propose a statistical approach to weighting estimations for data with an ordinal or binary outcome variable. Applying reciprocals of the sampling probabilities at each sampling stage to weight the log-likelihood function and testing the results through a simulation study, they show that the extension of the PML method to a binary or ordinal outcome variable

can yield consistent estimates of parameters. These studies are relatively recent, suggesting that the implications of informative sampling designs need to be further explored.

Time to event data collected through panel surveys often involve interval censored data. When the exact time to event of interest is not observed by a researcher, the event times are said to be censored. Specifically, event times are right-censored if study participants are lost to follow-up or do not experience the event by the end of the study, left-censored if they already experienced the event by the first observation time point, and interval-censored if they experience the event between two observation points and the exact time is not known. While right-censored event times can be accounted for by both Accelerated Failure Time (AFT) and Proportional Hazard (PH) models, interval-censored event times require model specification within the framework of AFT model (Allison 1995).

The standard survival analysis assumes that every individual is at risk for the event of interest, of which assumption may not hold when there is a subpopulation not at risk. Often surveys have a fixed period of time for observation or data collection. Individuals who do not realize an event by the end of study are recorded as right-censored observations. Those right-censored observations may be comprised of two different subpopulations: at-risk individuals whose survival time exceeded the last observation time point, and "long-term survivors" not at risk for the event. When the subpopulation not at risk for the event is not known to the researcher, the subpopulation cannot be distinguished from at-risk individuals whose survival time exceeded the study period. As a result, they are examined as at-risk individuals in the standard survival analysis, based on the assumption that they will eventually realize the event. Consequences of the inappropriate assumption include biased estimates of parameters (Taylor et al. 2003). Examples of such data can be found in the studies of time to conception when the population contains an unknown proportion of sterile subjects. In the example discussed in this paper we assume that there is an unknown proportion of subjects who can be classified as long-term survivors of obesity.

Previous research adopted mixture distribution models often involving two components: a survival distribution for at-risk individuals and a probability distribution of risk for the event (most commonly a Bernoulli distribution) (Berkson and Gage 1952; Boag 1949; Farewell 1982; Simon E. Pack and Byron J. T. Morgan 1990; Taylor et al. 2003). The early work was motivated by studies including, but not limited to, those by Boag (1949) and Berkson and Cage (1952). Farewell (1982) reviewed a mixture distribution model incorporating a Bernoulli distribution in a

parametric survival analysis model. Pack and Morgan (1990) extended a mixture distribution model to account for interval-censored event times. More recently, Taylor *et al.* (2004) applied a mixture distribution model incorporating a Weibull survival distribution and a Bernoulli distribution to sexually transmitted diseases data. However, studies on mixture distribution models are still few and require further research.

In summary, the three methodological challenges, i.e., clustered data with informative sampling probabilities, interval-censored time events, and an unknown subpopulation not at risk, suggest that parameter estimates may be biased and hypothesis testing results may be invalid if statistical analysis fails to account for them. In contrast to the increasing application of survival analysis to survey data in various fields in the last few decades, little research exists to address the problems. This study demonstrates a statistical approach to addressing the three challenges. First we use non-parametric maximum likelihood methods using Turnbull algorithm to estimate the Kaplan-Meier analog of survival function. We further extend the Pseudo Maximum Likelihood method to a random effect, mixture distribution model within the framework of Accelerated Failure Time model. Particular attention is paid to the scaling of sample weights in multistage sampling. We consider the National Longitudinal Study of Adolescent Health, conducted by the Carolina Population Center, University of North Carolina at Chapel Hill as an example and focus on time to obesity among adolescents and its association with major demographic characteristics including gender and race/ethnicity.

2. MOTIVATING EXAMPLE

Obesity among adolescents has become a more pressing public health problem in recent years in the United States. It is estimated that the proportion of obese adolescents (i.e., BMI equal to or greater than the 95th percentile of the sex-specific BMI growth charts) ages 12-19 increased from 5.0% to 18.1% between 1976-1980 and 2007-2008 (Ogden and Carroll 2010). Obesity has negative consequences for physical, psychological, and social well-being throughout the life course, including increased risk for adulthood obesity (Dietz 1998; Goodman et al. 2000; Harris et al. 2009; Reilly et al. 2003; Serdula et al. 1993; Whitaker et al. 1997), comorbidities (Dietz 1998; Harris et al. 2009), and deteriorated social life (Goodman et al. 2000; Gortmaker et al. 1993; Harris et al. 2009; Strauss and Pollack 2003). Adolescence obesity is associated with disparities among socio-demographic subgroups including gender and race/ethnicity (Goodman et al. 2000; Harris et al. 2009; Ogden and Carroll 2010; The et al. 2010), in which males and

racial/ethnic minorities are more likely to be overweight or obese than their respective counterparts. Research on time to obesity among adolescents is required to inform policies as to the timing and target population of interventions. We use the Add Health data to address the question.

Add Health is a nationally representative, school-based longitudinal study of adolescents aimed at exploring health related behaviors (Gordon-Larsen et al. 2004). This study focuses on adolescents interviewed through an in-home questionnaire which collects information from sampled students in grades 7 to 12 enrolled in selected schools. The first wave was carried out in 1994 -1995 (when the respondents are around the ages 12-20), followed by wave II in 1996 (when respondents are around the ages 13-21), wave III in 2001-2002 (when the respondents are around the ages 18- 26), and wave IV in 2007-2009 (when the respondents are between age 24-32.

At wave I, schools and adolescents were sampled based on a multistage sampling design. At the first stage, a nationally representative sample of 80 high schools and 52 middle schools was selected with probability proportionate to size. At the second stage, 20,745 adolescents in grades 7 through 12 enrolled in the sampled schools were selected with unequal sampling probabilities to allow oversampling of several ethnic-minority groups. Therefore, the Primary Sampling Unit (PSU) is school and the Secondary Sampling Unit (SSU) is individual adolescent. At wave II, 14,738 adolescents comprising a subsample of adolescents interviewed at wave I were followed up. Waves III and IV targeted all wave I respondents and followed up 15,197 and 15,701 adolescents and young adults, respectively. The analytical sample used in this study excludes respondents with missing data and is comprised of 18,466 respondents from 130 schools. The outcome is time to obesity, measured by respondent's age in years at the time of the onset of obesity, measured by body mass index (BMI). BMI values of 30 or greater or age- and sex-adjusted BMI percentiles of 95 or greater are classified as obese.

Because onset of obesity is observed only at each of the waves, the exact time to onset of obesity is unknown in this data. Instead we know that the observation to time to event is left censored (if the subject is already obese at wave I), interval censored if the subject is observed to be obese in between two waves and the right censored if the subject is not obese at last observation.

Descriptive statistics of the respondents, weighed by their final sampling probability, are presented in Table 1. Males and females are distributed almost evenly (51.0% and 49.0%,

respectively). The majority of respondents self-classified their race/ethnicity as non-Hispanic Whites (72.7%), followed by African-Americans (16.0%). Of the 18,466 respondents, 2,059 (11%) had left-censored event times, 12,549 (68%) had right-censored event times, and 3,848 (21%) had interval-censored event times.

Table 1. Descriptive statistics of respondents

	%
Gender	
Male	51
Female	49
Race/Ethnicity	
Non-Hispanic White	72.6
Other	27.4

Note: observations are weighted based on the final sampling probability

3. METHODS

Overview

We first estimate the survival curves using a non-parametric method proposed by Turnbull by incorporating sampling weights. We further extend the PML method to a random effect, mixture distribution model within the framework of AFT model to assess time to obesity among adolescents using the Add Health data. The model addresses the survey's potentially informative sampling design, clustered data, censored event times, and an unknown subpopulation not at risk for obesity. For all the analyses, SAS version 9.2 (SAS Institute, Cary, NC) is used.

Non-Parametric Estimation of Survival Curves

Let T_k denote the time to obesity for the k th respondents, for k =1, 2,… n. Let $(O_k, U_k]$ be the interval for which T_k is measured where O_{ij} and U_{ij} are lower and upper time points, respectively. Note that the interval $(O_k, U_k]$ may overlap across respondents. From the data of

overlapping intervals $(O_k, U_k]$, Turnbull (1976) derived a procedure to generate a set of non-overlapping intervals $\{q_l, p_l\} = \{(q_1, p_1], ..., (q_m, p_m]\}$ for $l = 1, ..., m$, over which the survival curve S(t) = $\mathbf{Pr}(T > t)$ can be estimated (Turnbull 1976). The set of non-overlapping intervals are obtained from all left, interval and right censored intervals in such a way that q_l is a left end point and p_l is the right end point, and there is no other left or right end point between q_l and p_l.

Turnbull (1976) proposed an EM algorithm to obtain a probability distribution (as well as the survival function). The procedure can be briefly summarized in the following way. Let $\theta_l = \Pr(q_l < T < p_l, l = 1, ..., m)$. Note that $\sum_{l=1}^{m} \theta_l = 1$. The iterative estimation of the parameters θ_l starts with initial estimates, usually assumed to be uniform in the intervals.

The revised estimates are calculated as:

$$\hat{\theta}_k(t_l) = \frac{\hat{\theta}(t_l) I\{t_l \in (O_k, U_k]\}}{\sum_{t_l \in (O_k, U_k)} \hat{\theta}(t_k)}.$$

Then in the maximization step, an improved estimate can be obtained as $\hat{\theta}(t_l) = \frac{1}{n} \sum_{k=1}^{n} \hat{\theta}_k(t_l)$ and the process is continued until convergence. In order to account for sampling weights, a weighted average is calculated in the maximization step.

Regression Analysis of Interval-Censored Event Times with Clustering

Suppose time to obesity of respondent j ($j = 1, ..., n_i$) from ith school ($i = 1, ..., N$) is only known to be in the interval $(O_{ij}, U_{ij}]$, where O_{ij} and U_{ij} are lower and upper time points (measured by respondent's age in years), respectively. That is, even time is right-censored if (O_{ij}, ∞), left-censored if $(0, U_{ij}]$, and interval-censored if both O_{ij} and U_{ij} are observed and $O_{ij} \neq U_{ij}$, and exactly observed (i.e., non-censored) if both O_{ij} and U_{ij} are observed and $O_{ij} = U_{ij}$. To account for clustering of observations within schools, we incorporate a frailty term

in the model. For computational simplicity we assume the frailties follows a normal (0, θ) distribution.

We apply a likelihood approach in estimating the model parameters. Suppose observations from school i are independent conditional on an unobserved frailty (b_i). For simplicity, we assume that time to obesity follows a Weibull (λ, p) distribution. Let $\mathbf{x}_{ija}$ be a vector of covariates (e.g., race/ethnicity and gender) for j th respondent from i th school, $\boldsymbol{\alpha}$ be a vector of corresponding coefficients, and $\lambda = \exp(-\alpha_0)$ .Then the individual likelihood can be written as:

$$L_{ij}(\boldsymbol{\alpha}, \theta, p) = \begin{cases} \exp\left\{-\left[O_{ij}\exp(-\mathbf{x}_{ij1}\boldsymbol{\alpha}) + b_i\right]^{1/p}\right\}, \text{ if } U_{ij} = \infty, \\ 1 - \exp\left\{-\left[U_{ij}\exp(-\mathbf{x}_{ij1}\boldsymbol{\alpha}) + b_i\right]^{1/p}\right\}, \text{ if } O_{ij} = 0, \\ \exp\left\{-\left[O_{ij}\exp(-\mathbf{x}_{ij1}\boldsymbol{\alpha}) + b_i\right]^{1/p}\right\} - \exp\left\{-\left[U_{ij}\exp(-\mathbf{x}_{ij1}\boldsymbol{\alpha}) + b_i\right]^{1/p}\right\}, \text{if } O_{ij} \neq U_{ij}, \\ \lambda p(\lambda O_{ij})^{(p-1)}\exp(-\mathbf{x}_{ij1}\boldsymbol{\alpha} + b_i)^{1/p}, \text{if } O_{ij} = U_{ij}. \end{cases} \quad (1)$$

Regression Analysis Involving Informative Sampling Weights

Suppose that an entire population (as opposed to a sample drawn from the population) contains M clusters (schools) with N_i subjects (respondents within school). Based on the information on the entire population containing M schools, the census log likelihood is specified as follows:

$$\log L(\boldsymbol{\alpha}, \theta, p) = \sum_{i=1}^{M} \log \int \left\{ \exp\left[\sum_{j=1}^{Ni} \log L_{ij}(\boldsymbol{\alpha}, \theta, p \mid b_i) \right] \right\} \phi(b_i) db_i \ , \quad (2)$$

However, the census likelihood cannot be obtained from a sample. Therefore the population quantities in the expression (2) are replaced by the analogous estimates based on the sample incorporating the sampling weights at the cluster and elementary observation unit levels.

Suppose the clusters and subjects are sampled as follows: At the first stage N clusters (schools) are selected with the inclusion probability π_i , $i = 1,..., N$. At the second stage n_i subjects (respondents) are selected within i th selected cluster with the inclusion probability $\pi_{j|i}\,(i = 1,..., n_i)$. Define the first stage sampling weight $w_i = 1/\pi_i$ and the second stage

sampling weight $w_{j|i} = 1/\pi_{j|i}$, respectively. Then a design-consistent estimate of the census likelihood can be obtained by incorporating the sampling weights at the first and second stages:

$$\log \hat{L}(\boldsymbol{\alpha}, \theta, p) = \sum_{i=1}^{N} w_i \log \int \left\{ \exp\left[\sum_{j=1}^{ni} w_{j|i} \log L_{ij}(\boldsymbol{\alpha}, \theta, p \mid b_i) \right] \right\} \phi(b_i) db_i \ . \qquad (3)$$

Similarly one can write the population score function $U(\theta) = \partial / \partial\theta \log L(\theta)$ and the corresponding sample estimate ($\hat{U}(\theta)$). Note that the implementation of the PML method using the sample quantities requires sampling weights at both levels. The sample score functions do not yield a closed from solution. Therefore iterative procedures are used to obtain parameter estimates. Bellamy *et al*. propose to approximate it through Gaussian quadrature.

Scaling of Weights

As noted above, sampling weights at both first and second sampling stages need to be introduced to generalize the PML method to a model accounting for an informative sampling design. We assume that w_i, the sampling probability at the first sampling stage, may be correlated with time to obesity. However, we assume that $w_{j|i}$, the conditional sampling probability at the second sampling stage is not correlated with time to obesity, after accounting for gender and race/ethnicity. These assumptions imply that the study design may be informative due to its sampling scheme at the first sampling stage.

One strategy to reduce bias in parameter estimates is to scale the weights. Scaling of the second stage sampling weight may have important effect on the small sample behavior of the PML estimator (Grilli and Pratesi 2004). Pfeffermann (1998) propose several approaches to scaling the conditional sampling weight ($w_{j|i}$) under different assumptions: 1) $w^*_{j|i} = (w_{j|i} \sum_j w_{j|i}) / \sum_j w^2_{j|i}$, when the study design is informative at both the first and second stages, and 2) $w^*_{j|i} = (w_{j|i} \times n_i) / \sum_j w_{j|i}$, when the study design is informative at the first stage.

Under the assumption that the study design is informative at the first stage, we apply the second method, i.e., $w^*_{j|i} = (w_{j|i} \times n_i) / \sum_j w_{j|i}$.

Regression Analysis Involving an Unknown Subpopulation Not at Risk

Finally, we assume that respondents with right-censored event times consist of two groups: adolescents at risk for obesity but did not become obese by the last observation point, and those not at risk for obesity. Let Y be a random variable representing risk for obesity. That is, Y_{ij} =1 if respondent j from school i is at risk and Y_{ij} =0 otherwise. Then the probability of being at risk can be written as (Taylor et al. 2003):

$$\Pr\{Y_{ij} = 1 \mid \mathbf{x}_{ij2}\} = \exp(\mathbf{x}_{ij2}\boldsymbol{\beta}) / \{1 + \exp(\mathbf{x}_{ij2}\boldsymbol{\beta})\}, \quad (4)$$

where $\mathbf{x}_{ij2}$ is a vector of covariates, which may be different from $\mathbf{x}_{ij1}$ (the vector of covariates associated with time to obesity) and $\boldsymbol{\beta}$ is a vector of corresponding coefficients. Suppose there are n_{i1} respondents whose event times were not right-censored and n_{i2} respondents whose event times were right-censored from school i, for $n_i = n_{i1} + n_{i2}$. Finally, introducing the mixture distribution to the function (3), the estimate of census likelihood incorporating the sampling design can be specified as follows:

$$\log \hat{L}(\boldsymbol{\alpha}, \theta, p) = \sum_{i=1}^{N} w_i \log \int \left\{ \exp \left[\begin{array}{l} \sum_{j=1}^{ni1} w^*_{j|i} \log\left(\Pr(Y_{ij} = 1 \mid \mathbf{x}_{ij2}) L_{ij}(\boldsymbol{\alpha}, \theta, p \mid b_i)\right) + \\ \sum_{j=1}^{ni2} w^*_{j|i} \log\left(\Pr(Y_{ij} = 0 \mid \mathbf{x}_{ij2}) + \left(\Pr(Y_{ij} = 1 \mid \mathbf{x}_{ij2}) L_{ij}(\boldsymbol{\alpha}, \theta, p \mid b_i)\right)\right) \end{array} \right] \right\} \phi(b_i) db_i \quad (5)$$

where $\Pr(Y_{ij} = 0 \mid \mathbf{x}_{ij2}) = 1 - \Pr(Y_{ij} = 1 \mid \mathbf{x}_{ij2})$. The estimated census likelihood function (5) accounts for the three aforementioned methodological challenges: clustered data with informative sampling weights, interval-censored event times, and an unknown subpopulation not at risk.

In this paper, we compare three regression models to examine the effects of the tree methodological challenges on parameter estimates and hypothesis testing. The first model is an AFT model ignoring clustering and informative sampling weights, based on the assumption that every individual is at risk (hereafter referred to as “unadjusted model”). The model incorporates the final sampling weight w_{ij}, which is a common practice when using survey data. Parameter

estimates are obtained through LIFEREG procedure. The second model is an AFT model incorporating frailty and informative sampling weights, based on the assumption that every individual is at risk (hereafter referred to as “random effect model”). Parameter estimates are obtained through NLMIXED procedure. The third model is an AFT model incorporating frailty and informative sampling weights, based on the assumption that there is a subpopulation not at risk (hereafter referred to as “mixture distribution model”). Parameter estimates are obtained through NLMIXED procedure.

The NLMIXED procedure of SAS is well suited for the analysis purposes because it can easily incorporate these model specifications. The weighting method is introduced via the statement REPLICATE in NLMIXED. Because REPLICATE statement allows only an integer, the sampling weight at the school level, w_i, is inflated by a constant (e.g., 100,000) and controlled by the statement CFACTOR (see Grilli and Prastesi (2004) for details). As discussed in a later section, the standard errors are obtained through a Jackknife procedure for the random effect model and the mixture distribution model.

The three models together can be considered a progression in addressing statistical assumptions. That is, the random effect model is built on the unadjusted model to address clustered data with informative sampling weights, and the mixture distribution model is built on the random effect model to address an unknown subpopulation not at risk. Differences in the parameter estimates, therefore, can be interpreted as the effect of addressing a statistical assumption that was ignored in the basis model.

Estimation of Variances

Because the aforementioned NLMIXED procedure involves replication, variances of parameters are not correctly estimated. Specifically, variances are underestimated due to the inflation of sample size. Therefore they need to be estimated through an alternative procedure.

The robust sandwich estimator as proposed by Skinner (1989) is ideal because it can provide an appropriate variance of parameter estimates. However, the proposed sandwich estimator is derived from single-level models and its computation is not straightforward to derive from multilevel models. Therefore we estimate variances through a Jackknife procedure. Let τ be a

parameter of interest. Korn and Grauard (1999) propose estimating its variance ($V(\tau)$) through a Jackknife procedure as follows (page 29):

$$\hat{V}(\hat{\tau}) = \frac{N-1}{N}\sum_{i=1}^{N}(\hat{\tau}_{(i)} - \hat{\tau})^2 \qquad (6)$$

where $\hat{\tau}$ is an estimate of τ based on the entire sample, and $\hat{\tau}_{(i)}$ is an estimate based on data excluding school i.

4. RESULTS

Figure 1 plots the survival curve (i.e., the cumulative proportion of those not obese) for all the respondents using the non-parametric procedure proposed by Turnbull (1976). The estimated cumulative probability of obesity by age 33 is 0.47.

Figures 2 and 3 plot survival curves stratified by race/ethnicity (i.e., non-Hispanic Whites vs. others) and gender (males vs. females), respectively. Non-Hispanic Whites have a higher survival probability until their early 30's, after which the gap in the survival probabilities closes between non-Hispanic Whites and non-Whites (Figure 2). The estimated probability of obesity by age 33 is 0.51 for non-Hispanic Whites and 0.49 for non-Whites, respectively. Males and females exhibit similar survival curves during adolescences and early adulthoods. The estimated probability of obesity by age 33 is 0.47 for both males and females.

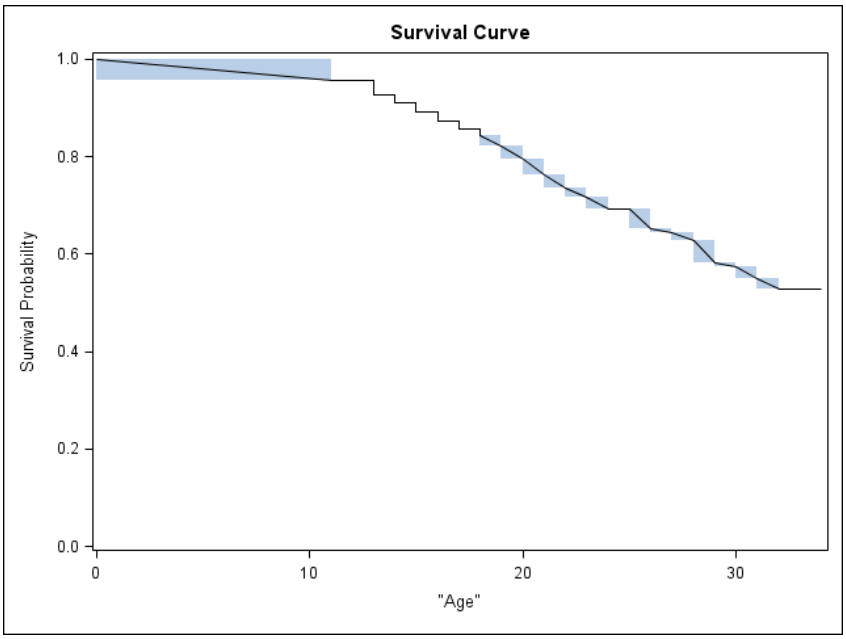

Figure 1. Survival curve of all the respondents

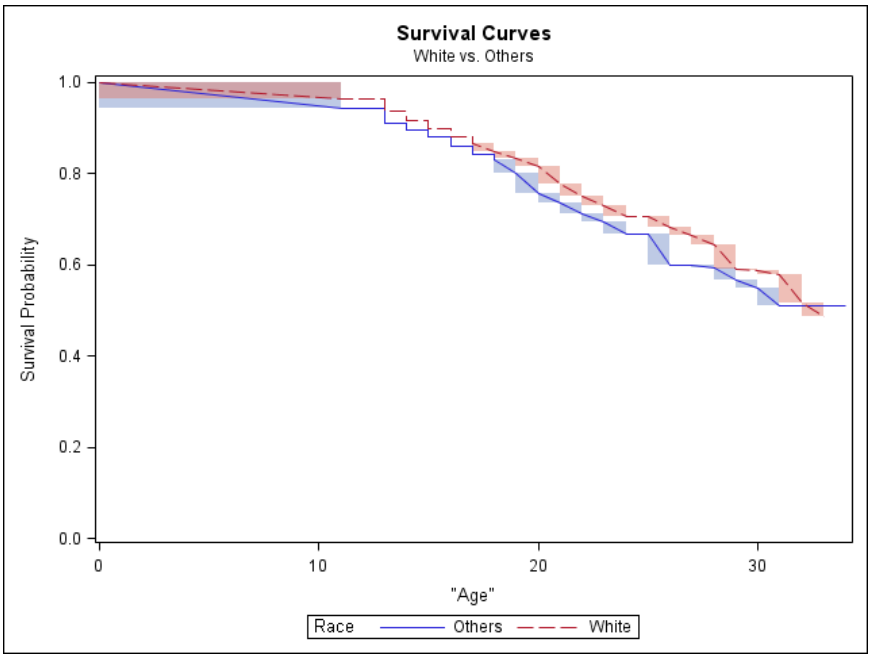

Figure 2. Survival curve by race/ethnicity

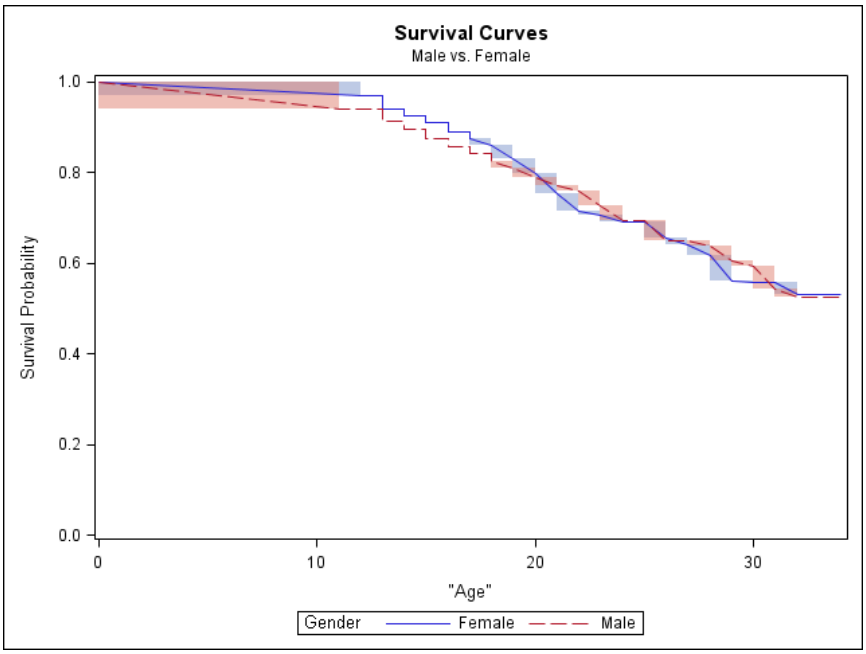

Figure 3. Survival curve by gender

The regression analysis results are presented in Table 2 for the three models: unadjusted model, random effect model, and mixture distribution model.

Table 2. Survival analysis results

	Unadjusted model (LIFEREG procedure)			Random effect model (NLMIXED procedure)			Mixture distribution model (NLMIXED procedure)		
	Coefficient estimate	(SE)	Pr(Z>|z|)	Coefficient estimate	(SE)	Pr(Z>|z|)	Coefficient estimate	(SE)	Pr(Z>|z|)
Weibull									
Intercept	3.533	0.0189	<.001	3.569	0.0258	<.001	3.278	0.0098	<.001
White	0.148	0.0196	<.001	0.105	0.0303	<.001	0.063	0.0121	<.001
Males	0.109	0.0162	<.001	0.063	0.0191	<.001	-0.130	0.0087	<.001
Interaction	-0.116	0.0192	<.001	-0.087	0.0249	<.001	-0.032	0.0162	0.048
Bernoulli									
Intercept	-			-			0.971	0.0195	<.001
White	-			-			-0.390	0.0231	<.001
Males	-			-			-1.013	0.0269	<.001
Interaction	-			-			0.396	0.0271	<.001
Scale	0.447	0.0091		0.43	0.0093		0.356	0.0100	
Theta	-			0.012			0.016		

Results from the Unadjusted Model

The estimated average time to obesity obtained through the unadjusted model is 34.2 years (i.e., $\exp(3.533)$) among non-White females. It is 16.0% (i.e., $100\times\{\exp(0.148)-1\}$) longer among non-Hispanic White females and 11.5% (i.e., $100\times\{\exp(0.109)-1\}$) longer among non-White males than among non-White females. Overall, non-Hispanic White males have an average time to obesity 15.1% longer than non-White females. These coefficients are significantly different from 0 at the two-side α level of .05, based on the standard errors obtained through LIFEREG procedure. The corresponding hazard of obesity is 72 % (of the hazard for non-White females) for non-Hispanic White females; and 78 % (of the hazard for non-White females) for non-White males is 78 %, and 73 % (of the hazard for non-White females) for non-Hispanic White males.

Results from the Random Effect Model

The estimated average time to obesity obtained through the random effect model is 35.5 years (i.e., $\exp(3.569)$) among non-White females, which is slightly larger compared with the unadjusted model. The estimates of the other regression coefficients are closer to the null compared with the unadjusted model, suggesting that the coefficients are overestimated in the unadjusted model as a result of ignoring informative sampling weights. The differences, however, are relatively small and do not change the conclusions as to the association between the studied socio-demographic factors and time to obesity. The standard errors obtained through a Jackknife procedure are larger compared with the unadjusted model, suggesting that the unadjusted model underestimates the parameter variances by ignoring clustering and informative sampling weights.

Results from the Mixture Distribution Model

On the other hand, the results of the mixture distribution model are distinct from the previous two models. The estimated time to obesity among non-White females at risk for obesity is 26.5 years (i.e., $\exp(3.278)$), which is shorter than the previous two models. The difference in time to obesity between non-Hispanic Whites and non-Whites at risk for obesity is also smaller compared with the other two models. The study conclusion as to the association between gender and time to obesity is reversed, given the negative coefficient estimate of -0.130; the average time to obesity is 12.2% shorter among non-White males than among non-White

females at risk for obesity. Overall, the average time to obesity is 9.4% shorter among non-Hispanic White males at risk than among other non-White females at risk.

At the same time, the results suggest that there may be a subpopulation not at risk for obesity and that non-Hispanic Whites and males have lower risk for obesity, compared with non-Whites and females, respectively. The estimated probability of obesity among non-Hispanic White females is 0.641 (i.e., $1-1/\{1+\exp(0.971-0.390)\}$), while that among non-White females is 0.725 (i.e., $1-1/\{1+\exp(0.971)\}$). Likewise, the estimated probability of obesity among non-White males is 0.490 (i.e., $1-1/\{1+\exp(0.971-1.013)\}$). Overall, non-Hispanic White males have a probability of 0.491 (i.e., $1-1/\{1+\exp(0.971-0.391-1.013+0.396)\}$).

Summary

We confirm that the magnitude of regression coefficients can be biased when informative sampling design is not accounted for. Also the variances are underestimated when clustering is ignored. We also confirm that the covariates (i.e., socio-demographic factors) may exhibit a different association with time to obesity and the probability of obesity. When the two different subpopulations, one at risk and the other not at risk for obesity, are not addressed in a model, the regression coefficients may be biased and mask the true association between covariates and time to obesity.

5. DISCUSSION

Survival analysis may face methodological challenges when analyzing complex survey data: clustered data with informative sampling weights, interval-censored event times, and an unknown subpopulation not at risk. This study addresses the three challenges by extending the PML method to a random effect, mixture distribution model within the framework of AFT model. First, the likelihood function accounted for interval-censored event times. Second, based on the assumption that the study design is informative at the school level, we introduced weights at the school and individual levels to the likelihood function. At the same time, a frailty term was incorporated to account for clustering. Third, a mixture distribution model was introduced to account for an unknown subpopulation not at risk for obesity. This study therefore presents a statistical approach to eliminating bias due to multistage, informative sampling designs, an unknown subpopulation not at risk using, and multiple types of censoring.

The study results confirm that a model failing to address the problems may produce biased estimates of parameters. It is therefore critical to apply a method that best addresses the sampling design and characteristics of the study population. Specifically, the magnitude of regression coefficients can be biased when informative sampling design is not accounted for. Also the variances are underestimated when clustering is ignored. The bias due to informative sampling design and clustering in this study is not as large as to reverse the study conclusions. The unadjusted model and the random effect model assuming risk for obesity for the entire population identify non-Whites and females as having a shorter time to obesity.

On the other hand, the mixture distribution model provides different conclusions especially as to the association of covariates with time to obesity and probability of obesity. Non-Hispanic White adolescents are found to have a longer time to obesity and lower risk for obesity compared with their non-White counterparts. On the other hand, males have a shorter time to obesity when at risk for obesity but have lower risk for obesity compared with females. They may imply that there are more females than males at risk for obesity, but males at risk for obesity are likely to become obese earlier than females at risk. When the two different subpopulations, one at risk and the other not at risk for obesity, are not addressed in a model, the regression coefficients may be biased and mask the true association between covariates and time to obesity. The study results therefore suggest that further research is required to address time to event and probability of event separately.

This study has several limitations. First, the anthropometric information collected at wave I is based on self-reporting, unlike that at waves II, III, and IV, which is measured by trained interviewers. Therefore it is possible that the anthropometric information at wave I contains measurement error. To examine measurement error of self-reported weight and weight among adolescents and its impact on study results, Goodman *et al*. (2000) compare obesity status between waves I and II, using the Add Health data. They conclude that 96% of respondents are correctly classified as to obesity status when self-reported weight and height are used to calculate BMI. Because they assess the same data used in this study, their conclusion may be generalized to this study, suggesting that the impact of measurement error on the study results may be minimal, if any at all.

Second, the probability estimation is available only for Turnbull intervals. While the estimation method is well suited for the Add Health data because the respondents were observed in

overlapping intervals, the method may not be applicable to other data if the intervals are fixed or non-overlapping. Therefore, exploration of data and careful application of the method is required.

Future Work

Further work is planned and will be included in the final draft of this paper to be uploaded on the PAA 2012 website by April 2, 2012.

First, we plan to refine the mixture distribution model. In this preliminary report, we assumed that both gender and race/ethnicity are associated with both time to obesity and risk for obesity. However, it is possible that different sets of covariates may be associated with the two outcomes. Future work therefore includes exploration and selection of covariates into the different components of distributions, i.e., Weibull and Bernoulli distributions, through AIC and BIC.

Second, we plan to estimate parameter variances through bootstrapping. In this preliminary report, we used a Jackknife procedure to estimate parameter variances because the ideal robust sandwich estimator is computationally intense to obtain in the multilevel models. However, the procedure is known not to be optimal; in addition, performance of Jackknife estimates in presence of informative sampling designs is not well studied (Grilli and Pratesi 2004). Given the concern, Grilli and Pratesi (2004) recommend estimation through bootstrapping to obtain the variance covariance matrix. Future work therefore includes estimation of variances through bootstrapping.

ACKNOWLEDGEMENTS

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

REFERENCES

Allison, P.D. 1995. *Survival analysis using SAS : a practical guide*. Cary, NC: SAS Institute.

Berkson, J. and R.P. Gage. 1952. "Survival Curve for Cancer Patients Following Treatment." *Journal of the American Statistical Association* 47(259):pp. 501-515.

Boag, J.W. 1949. "Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy." *Journal of the Royal Statistical Society.Series B (Methodological)* 11(1):pp. 15-53.

Dietz, W.H. 1998. "Health Consequences of Obesity in Youth: Childhood Predictors of Adult Disease." *Pediatrics* 101:518-525.

Farewell, V.T. 1982. "The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors." *Biometrics* 38(4):pp. 1041-1046.

Goodman, E., B.R. Hinden, and S. Khandelwal. 2000. "Accuracy of Teen and Parental Reports of Obesity and Body Mass Index " *Pediatrics (Evanston)* 106(1):52-58.

Gordon-Larsen, P., L.S. Adair, M.C. Nelson, and B.M. Popkin. 2004. "Five-year obesity incidence in the transition period between adolescence and adulthood: the National Longitudinal Study of Adolescent Health " *The American Journal of Clinical Nutrition* 80(3):569.

Gortmaker, S.L., A. Must, J.M. Perrin, A.M. Sobol, and W.H. Dietz. 1993. "Social and Economic Consequences of Overweight in Adolescence and Young Adulthood " *The New England Journal of Medicine* 329(14):1008-1012.

Grilli, L. and M. Pratesi. 2004. "Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs." *Survey Methodology* 30(1):93-103.

Harris, K.M., K.M. Perreira, and D. Lee. 2009. "Obesity in the Transition to Adulthood: Predictions Across Race/Ethnicity, Immigrant Generation, and Sex " *Archives of Pediatrics & Adolescent Medicine* 163(11):1022-1028.

Kish, L. 1965. *Survey sampling*. New York: Wiley 1995.

Korn, E.L. and B.I. Graubard. 2003. "Estimating Variance Components by Using Survey Data." *Journal of the Royal Statistical Society.Series B (Statistical Methodology)* 65(1):pp. 175-190.

Lohr, S. 1999. *Sampling : design and analysis*. Pacific Grove, CA: Duxbury Press.

Ogden, C.L. and M.D. Carroll. 2010. "Prevalence of Overweight, Obesity, and Extreme Obesity Among Adults: United States, Trends 1976–1980 Through 2007–2008 " .

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for unequal selection probabilities in multilevel models " *Journal of the Royal Statistical Society.Series B, Statistical Methodology* 60(1):23-40.

Reilly, J.J., E. Methven, Z.C. McDowell, B. Hacking, D. Alexander, L. Stewart, and C.J.H. Kelnar. 2003. "Health consequences of obesity " *Archives of Disease in Childhood* 88(9):748-752.

Serdula, M.K., D. Ivery, R.J. Coates, D.S. Freedman, D.F. Williamson, and T. Byers. 1993. "Do obese children become obese adults? A review of the literature " *Preventive Medicine* 22(2):167.

Simon E. Pack and Byron J. T. Morgan. 1990. "A Mixture Model for Interval-Censored Time-to-Response Quantal Assay Data." *Biometrics* 46(3):pp. 749-757.

Skinner, C.J. 1989. "Domain means, regression and multivariate analysis " in *Analysis of Complex Surveys* , edited by C.J. Skinner, D. Holt, and T.M.F. Smith. Chichester: Wiley.

Strauss, R.S. and H.A. Pollack. 2003. "Social Marginalization of Overweight Children " *Archives of Pediatrics & Adolescent Medicine* 157(8):746-752.

Taylor, D.J., M.A. Weaver, and R.E. Roddy. 2003. "Evaluating factors associated with STD infection in a study with interval-censored event times and an unknown proportion of participants not at risk for disease " *Statistics in Medicine* 22(13):2191-2204.

The, N.C., C. Suchindran, K.E. North, B.M. Popkin, and P. Gordon-Larsen. 2010. "Association of Adolescent Obesity With Risk of Severe Obesity in Adulthood " *JAMA : The Journal of the American Medical Association* 304(18):2042-2047.

Turnbull, B.W. 1976. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data." *Journal of the Royal Statistical Society.Series B (Methodological)* 38(3):pp. 290-295.

Whitaker, R.C., J.A. Wright, M.S. Pepe, K.D. Seidel, and W.H. Dietz. 1997. "Predicting Obesity in Young Adulthood from Childhood and Parental Obesity " *The New England Journal of Medicine* 337(13):869-873.