

8-2-2008

# HIV/Aids Relative Survival and Mean Residual Life Analysis

Xinjian Zhang

Follow this and additional works at: [http://scholarworks.gsu.edu/math\\_theses](http://scholarworks.gsu.edu/math_theses)

---

## Recommended Citation

Zhang, Xinjian, "HIV/Aids Relative Survival and Mean Residual Life Analysis." Thesis, Georgia State University, 2008.  
[http://scholarworks.gsu.edu/math\\_theses/43](http://scholarworks.gsu.edu/math_theses/43)

This Thesis is brought to you for free and open access by the Department of Mathematics and Statistics at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Mathematics Theses by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact [scholarworks@gsu.edu](mailto:scholarworks@gsu.edu).

## **HIV/AIDS RELATIVE SURVIVAL ANALYSIS**

**BY**

**XINJIAN ZHANG**

**Under the Direction of Gengsheng ( Jeff ) Qin and Ruiguang (Rick) Song**

### **ABSTRACT**

Generalized linear models with Poisson error were applied to investigate HIV/AIDS relative survival. Relative excess risk of death within 3 years after HIV/AIDS diagnosis was significantly higher for non-Hispanic blacks, American Indians and Hispanics compared with non-Hispanic Whites. Excess hazard of death was also higher among male injection drug users compared with men who have sex with men (MSM). The relative excess hazard of old HIV/AIDS patients was significantly higher compared with younger patients (e.g., 60+ age group versus 19-29 year age group). When CD4 increased, the relative excess hazard decreased; while with the increase of HIV viral load, the relative excess hazard decreased. Our population-based results showed that viral load was a determinant risk factor of disease progression after HIV infection; basically the mean residual life has similar trend to relative survival.

**INDEX WORDS:** Human Immunodeficiency Virus (HIV), Acquired Immunodeficiency Syndrome (AIDS), Survival, Mean residual life (MRL).

**HIV/AIDS RELATIV ESURVIVAL ANALYSIS**

**BY**

**XINJIAN ZHANG**

Thesis submitted as Partial Fulfillment of Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2007

Copyright by  
XINJIAN ZHANG  
2007

**HIV/AIDS RELATIVE SURVIVAL ANALYSIS**

**BY**

**XINJIAN ZHANG**

Major Professor: Gengsheng (Jeff) Qin and  
Ruiguang (Rick) Song  
Committee: Xu Zhang  
Yu-sheng Hsu

Electronic Version Approved:

Office of Graduate Studies  
College of Arts and Sciences  
Georgia State University

August 2007

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following persons for their suggestion and assistance in the preparation of this thesis. Dr. Gengsheng (Jeff) Qin served as my thesis advisor, gave me invaluable guidance and constant encouragement. Without his direction and supervision, this thesis would not have been possible. Dr. Ruiguang Song, as my thesis field director, spent countless hours in shaping the direction of this thesis, offering me concrete suggestions in data manipulation and analysis. Many thanks to Dr. Yu-sheng Hsu for providing critical advices.

I am deeply grateful to Dr. Kathleen McDavid Harrison for her advices and expertise in HIV/AIDS epidemiology and surveillance. I also would like to thank the HIV Incidence and Case Surveillance Branch at Centers for Diseases Control and Prevention (CDC) for giving me the opportunity to access and analyze the HIV/AIDS reporting system (HARS) data. They paved the way for me to get this thesis done.

Finally, I would like to thank my family and my friends who gave me constant support and encouragement through the whole process.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER	
1 INTRODUCTION	1
History of HIV Infection	1
Epidemiology	2
Clinical Manifestation	3
Biology	5
Laboratory Diagnosis	6
Treatment and Prevention	7
Statistical Advance and Challenge in HIV/AIDS Modeling	9
HIV/AIDS relative survival and mean residual life analysis	10
2 METHODS	11
Data sources	11
Models and Estimation Approaches of Relative Survival and Mean Residual Life	13
Variable Selection and Model Checking	16

3	Data Analysis and Results	16
	Variable selection and survival curve analysis	17
	Relative survival from AIDS diagnosis to death	18
	Relative survival from HIV infection diagnosis to death	19
	Relative excess hazard ratio	20
	AIDS free survival proportion from HIV to AIDS diagnosis	21
	Model comparison and model checking	22
4	Discussion	22
5	Conclusions	28
	REFERENCES	29
	APPENDICES	52
	A    ABBREVIATIONS	52
	B    SAS AND SPLUS CODES	53



**LIST OF TABLES**

Table 1 Relative survival after AIDS diagnosis from 1996 to 2003, by 50 US States	34
Table 2 Relative survival after diagnosis from 1996 to 2003, by 25 US States	36
Tables 3 Relative excess hazard rate of death within 3 years among patients with HIV or AIDS diagnosis	38
Table 4. Relative survival from HIV to AIDS diagnosis during 1996 – 2003, by 25 US States	40
Table 5 Model comparison: Estimated excess hazard ratios and standard errors of the log excess hazard ratio	42

**LIST OF FIGURES**

Figure-1. Maximum likelihood tree of HIV	44
Figure-2. Survival analysis after AIDS diagnosis by race/ethnicity	45
Figure-3. Survival analysis after AIDS diagnosis by age group	46
Figure-4. Survival analysis after AIDS diagnosis by transmission category	47
Figure-5. Survival analysis after AIDS diagnosis by diagnosis year	48
Figure-6 Survival analysis after AIDS diagnosis by sex	49
Figure-7 Survival analysis after AIDS diagnosis by CD4 count	50
Figure-8 Survival analysis after AIDS diagnosis by HIV viral load	51

## CHAPTER 1. INTRODUCTION

### 1. History of HIV Infection

From late the 1970s to early 1980s in the United States and Europe, the emergence of immunologic dysfunction among patients of unknown etiology attracted popular attention [1]. Following the unusual occurrence of *Pneumocystis carinii* pneumonia (PCP) in five homosexual men from Los Angeles was reported to the Centers for Disease Control and Prevention (CDC) in 1981 [2], several similar reports, describing male homosexuals and intravenous drug users with impaired immune systems and T lymphocytes were sent to CDC [3]. Because more than 90% of these cases occurred in homosexual or bisexual men, the new term “gay-related immunodeficiency” (GRID) was coined. However, in 1982, hundreds of similar cases of AIDS had been reported, not only in homosexual and bisexual men, but also in intravenous drug users, hemophiliacs, blood-transfusion recipients, heterosexual adults from the Caribbean and Central Africa, sex partners of infected patients, and infants born to mothers with the syndrome. The common symptoms of the AIDS patients were due to the depletion of CD4<sup>+</sup> T-lymphocyte subset.

The similar epidemiological pattern implicated that a new emerging disease was transmitted by a novel pathogen in contaminated blood or fluid through sexual intercourse and other contact with an affected individual. In AIDS patients, the presence of lymphadenopathy in many affected individuals was reminiscent of the clinical course related to human viral pathogens, the attention was focused on viruses to infect cells of the immune system. Several candidate viruses were isolated, including cytomegaloviruses (CMVs), and human T-cell leukemia virus type-I (HTLV-I) [4]. In 1983, scientists at the

Pasteur Institute discovered a virus from the lymph nodes of an asymptomatic individual; the scientists presented their discovery with Gallo and colleagues at the National Institutes of Health [5]. Subsequently Gallo reported the isolation of retroviruses from AIDS patients, which they named HTLV-III. In 1994 Levy et al reported a similar retrovirus isolated from both AIDS patients and healthy individuals from the various risk groups [6] which they named the AIDS-associated retrovirus (ARV). The new retrovirus, associated with AIDS in the United States, Europe, and central Africa and exhibiting typical morphologic and genetic characteristics of the Lentivirus genus, was named human immunodeficiency virus (HIV) [7], and subsequently HIV-1. In 1986, a related, but immunologically distinct and less pathogenic human retrovirus (now called HIV-2), was recovered from individuals residing in several Western African countries [8].

## **2. Epidemiology**

The HIV pandemic is one of the most notorious infectious disease epidemics in human history. Its morbidity and mortality rates are staggering. In 2001, there were 36 million HIV-infected individuals worldwide [9]. In 2005, it was estimated that 38.6 (range 35~46) million people worldwide were living with HIV. More than 21.8 million deaths have been due to HIV infection since the beginning of the epidemic [9]. In 2005 alone, 3.4~4.2 million people died of AIDS and nearly 4.1 (range 3.4~6.2) million individuals worldwide acquired HIV infection [9]. There were an estimated 925,000-1,025,000 persons living with HIV/AIDS at the end of 2003 in the US., and approximately 40,000 new HIV infections occur each year [10]. HIV is transmitted by:

- mucosa contact (oral, rectal, or vaginal)during sex;
- transfusion of HIV contaminated blood products, use of contaminated equipment ; and
- maternal-fetal circulation or by breast feeding.

Sexual transmission accounts for more than 90% of HIV infections worldwide [10]. Transmission of HIV is dependent on behavioral and biologic factors. The probability of male-to-female HIV transmission during vaginal sex is approximately 0.1% to 0.2% per contact; receptive anal intercourse is associated with a considerably higher risk (0.82/per contact ) of HIV transmission [11]. Risk factors (variables) associated with transmission included the genetic background of the host, the size of the injured mucosa, and the local environment in which the exposure occurs.

With the advent of highly active antiretroviral therapy (HAART), the morbidity and mortality rates of HIV infection are decreasing dramatically in Europe and the USA [12]. However, in some of Africa and Asian countries, HIV morbidity and mortality rates are increasing due to the ineffective implementation of prevention and intervention policies [13].

### **3. Clinical Manifestation**

Usually, HIV infection includes a long period (approximately 10 years) of clinical latency between the time of primary infection and the development of symptoms indicative of advanced immunodeficiency. The clinical symptoms of HIV infection were evolved on the depletion of CD4<sup>+</sup> T-cell lymphocytes and the replication of HIV RNA. In the acute and early stages of the disease, there are some “flu-like” and “mononucleosis-like” illnesses, such as fever, rash, pharyngitis, and “lymphadenopathysis-like” illnesses.

Additionally, myalgias, arthralgias, diarrhea, nausea, vomiting, headache, hepatosplenomegaly, weight loss, thrush, and neurological symptoms may appear. For patients developing symptoms during primary HIV infection, the mean duration of symptoms is 3 weeks [14].

When CD4<sup>+</sup> T-cell count falls below 500 cells/uL, the following symptoms can be observed: oropharyngeal, recurrent vulvovaginal candidiasis, bacillary angiomatosis (usually due to infection with *Bartonella henselae*), recurrent or multidermatomal herpes zoster, listeriosis, infections due to *Rhodococcus equi*, pelvic inflammatory disease, oral hairy leukoplakia associated with EBV, cervical dysplasia (usually associated with human papillomavirus infection), constitutional symptoms such as unexplained fever or diarrhea lasting more than 1 month, idiopathic thrombocytopenic purpura, and peripheral neuropathy.

In late stage of HIV infection, when CD4<sup>+</sup> T-cell count is below the level of approximately 200 cells/uL, cellular immune responses are tremendously suppressed. Which cause many opportunistic infections, such as chronic microsporidiosis, gastrointestinal infection with *Cyclospora cayentanensis*, disseminated *Penicillium marneffeii* infection (endemic to southeast Asia), cerebral or disseminated *Trypanosoma cruzi* infection (endemic to Latin America), relapsing or chronic visceral leishmaniasis, anal carcinoma, and EBV-positive cases of leiomyosarcoma, leiomyosarcoma, and Hodgkin's disease. The increased risk in developing AIDS-defining illnesses associated with a CD4<sup>+</sup> T-cell count of less than 200 cells/uL led to the 1993 revision of the CDC definition of

AIDS, which includes a low CD4<sup>+</sup> T-cell count (<200 cells/uL) as an AIDS-defining criterion [15].

#### 4. Biology

Human immunodeficiency virus (HIV) is a member of the retrovirus family in the retroviridae (1). HIV is a genetically diverse population of viruses that is responsible for causing AIDS in much of West Africa (referred to as HIV-2), and causing AIDS throughout the rest of the world referenced as HIV-1 (1).

Based on the nucleotide acid sequence of complete viral genomes, HIV-1 was subsequently defined and classified into three groups: M (major), O (outlier), and N (non-M or O). The M group of HIV-1, which includes over 95% of the global virus isolates, consists of at least eight discrete clades A, B, C, D, F, G, H, and J (see the maximum likelihood tree in Figure 1).

The first step for HIV to infect a target cell is to bind the HIV receptor and coreceptors expressed on the surface of the target cells (such as T help lymphocytes, which are called as CD4<sup>+</sup> T-helper/inducer subset of lymphocytes, abbreviated as **CD4<sup>+</sup> T-cells or CD4**). Consequently, the main cellular targets for HIV-1 are the CD4<sup>+</sup> T-helper/inducer subset of lymphocytes, CD4<sup>+</sup> cells of macrophage lineage, and some populations of dendritic cells.

In order to quantitatively measure the amount of HIV in serum of HIV-infected patients, many technologies were developed, the commonly recognized method was polymerase chain reaction (PCR), the copy number of HIV RNA per micro liter serum

could be used as a rapid and sensitive measure of viral load, the copy number of HIV RNA in serum was often referred as **HIV viral load** [16].

HIV infection can cause the selective depletion of CD4<sup>+</sup> cells, therefore, **CD4<sup>+</sup> T-cell count** and **HIV viral load** are two of the most important variables to predict the clinical prognosis of HIV-infected subjects.

## 5. Laboratory Diagnosis

The specific biomarkers (antigen and antibodies) of HIV infection can be captured by serologic tests. Commonly, Enzyme-Linked ImmunoSorbent Assay (ELISA) was the preliminary screening test for HIV infection. Because ELISA often generates some false positive results, a positive ELISA test must always be followed by a confirmatory Western blot assay [17]. The ideal diagnosis marker of HIV infection should be easily and reproducibly measurable in all individuals with the disease. Furthermore, it should worsen with progression of disease and improve with positive responses to therapy.

CD4<sup>+</sup> T cell count is the “Marker of Immune System Dysfunction” [18]. CD4<sup>+</sup> T cells are the primary targets of HIV infection. Depletion of CD4<sup>+</sup> T cells is the immunologic hallmark of HIV disease progression. Measurement of CD4<sup>+</sup> T cells should be an excellent marker of disease progression. Some investigations have demonstrated that the CD4<sup>+</sup> T-cell count is a powerful predictor of the short-term risk of developing an AIDS-defining illness [19]. CD4<sup>+</sup> T-cell counts frequently increase in response to antiretroviral therapy, and this salutary response has been used as a criterion in the licensing of new anti-HIV drugs. A number of other indicators of immune dysregulation



correlate with HIV disease progression. The percentage of peripheral blood CD8<sup>+</sup> T cells bearing the activation marker CD38 is positively correlated with disease progression [20].

Measurement of the CD4<sup>+</sup> T-cell count yields information regarding the degree of immunodeficiency; quantification of HIV RNA in plasma obtains the information related to the rate and severity of immune deficiency. Technology currently available for reproducible measurement of plasma viremia in clinical specimens includes PCR, RT-PCR, nucleic acid sequence–based amplification (NASBA), and branched DNA (b-DNA) assays. The plasma viremia level of more than 100,000 HIV RNA copies/mL at 6 months after seroconversion was associated with an odds ratio of 10.8 for the development of AIDS. The level of plasma viremia at 7 to 12 months after seroconversion was a strong predictor of CD4<sup>+</sup> T-cell depletion during follow-up [1, 19].

Current recommendations regarding the initiation and maintenance of antiretroviral therapy rely heavily on the two best laboratory markers for HIV disease progression, the CD4<sup>+</sup> T-cell count and the HIV viral load [21-22].

## **6. Treatment and Prevention**

Highly active antiretroviral therapy (HAART), a single or combination of several drugs, has high activity to inhibit HIV RNA replication. HIV cocktail therapy is a combination reagent, which inhibits the replication of HIV RNA at different steps of HIV life-cycling. The currently available HIV inhibition reagents can be categorized into reverse transcriptase inhibitors, protease inhibitors, and integrated zinc-finger inhibitors.

Inhibitors of viral reverse transcriptase (RT) like zidovudine and lamivudine were the first agents to be developed for the treatment of HIV infection. The protease inhibitors, such as saquinavir, ritonavir, indinavir, nelfinavir, amprenavir, and lopinavir, are all potent antiretrovirals, particularly when used in combination with nucleoside analog RT inhibitors (NRTIs) [23].

Adherence to an antiretroviral treatment is a complex issue since antiretroviral regimens are multiple agents, and must be taken indefinitely in most cases. Many of the drugs have intolerable side effects and toxicities (nausea, vomiting, and/or diarrhea). Lapses in adherence provide the opportunity for virus replication, and enhance the probability of the emergence of drug-resistant virus mutants. Another insurmountable challenge for therapy is the presence of an HIV latent reservoir in resting CD4<sup>+</sup> memory T cells [24].

A treatment regimen should be selected that will afford a high likelihood of long-term profound suppression of plasma viral load and a significant increase in the CD4<sup>+</sup> T-cell count. Currently, the most useful prophylaxis strategy for HIV infection is to stop the virus transmission: condom use, blood screening, and preventing mother to child transmission.

An effective weapon in preventing further spread of the HIV epidemic will be a safe and effective vaccine. Even though tremendous efforts have been made and several promising candidate vaccines have been generated, there are huge unknown puzzles that remain to be elucidated in this field.

## **7. Statistical Advances and Challenges in HIV/AIDS Modeling**

Many statistical methodologies have been developed for modeling the HIV epidemic [25, 26], new therapies (HAART), prophylaxis strategies [27, 28], and HIV incidence [29-32]. Many of challenges remain to be elucidated by collaborative efforts from mathematicians, biologists and epidemiologists. For the treatment of HIV infection, the common philosophy is: “treat early and treat hard” [22, 23]. Because HIV genetic divergent mutants emerge constantly, the corresponding treatment and prevention methods need to be created accordingly. Therefore, it is indispensable to develop and identify optimal treatment strategies for HIV-infected individuals. How to determine an optimal treatment regimen is a great challenge. At present, there are a large number of treatment options, statisticians need to help clinicians figure out the following questions:

What is the best initial treatment combination?

When is the best time for a certain patient to switch from one regimen to another?

What multi-drug treatment sequence generates the best treatment effect?

Another statistical challenge is to estimate the efficacy of intended treatment. Since many potential treatment strategies become available, the reported promising new drugs may coincide with relevant clinical trials in progress. Thus, the patients may be attracted to withdraw from the trials to which they are party. The investigators may change the initial protocol to reflect the new achievement. Under this circumstance, the intend-to-treat rule may not generate unbiased estimation [33].

The most commonly used bio-markers, CD4<sup>+</sup> T-cell count and viral load, are subject to high variability during disease progression. It is necessary to measure them repeatedly over time for each patient. The treatment efficacy on the marker progress can be estimated by a standard multivariate model or mixed model. CD4<sup>+</sup> T cell count and HIV viral load often have missing values due to the death of patients or loss of follow up. Several non-parametric methods have been developed for solving the missing data problem [34, 35].

### **8. HIV/AIDS relative survival and mean residual life analysis**

Conventionally, net or crude survival was employed to analyze and predict patients' clinical disease progression after diagnosis. Recently, relative survival has become a very popular survival analysis method for cancer patients [36-38]. The prerequisite for net or cause-specific survival analysis is that fact of death information should be available. For HIV infected patients, it will generate some confusion when net survival is selected to extrapolate the survival of HIV-infected patients. According to the CDC 1993 definition, factors other than HIV infection can cause death of AIDS patients. HIV-infected patients have different opportunistic infections (OIs) as well as divergent co-infection (HBV, HCV, CMV, HPV, HSV). It is not a rare event that causes other than HIV infection will appear in the death certificates of AIDS patients. Therefore, it is more realistic to use relative survival to model and analyze disease progression of HIV/AIDS patients [38, 39]. For HIV/AIDS patients, relative survival is estimated as the ratio of observed survival of the HIV/AIDS patients (where all deaths are considered as events) to the expected survival of the general population, matched to the patients with respect to the main factors (age, race, sex, calendar year) and assumed to be free of HIV infection.

Mean residual life is a measurement of the remaining life expectancy of a subject at time  $t$ , which is the remaining survival time given the subject surviving up to  $t$ . It has been widely applied to many fields [40].

Since relative survival analysis can be used to identify the determinant factors of HIV/AIDS clinical progression, it has been taken as a reliable approach to estimate the efficacy of prophylaxis and treatment strategies of HIV-infected subjects. However, inconsistent results have been reported from different cohort studies. The inconsistency was related to the divergence of sample size, follow up time periods and the failure to correct important co-founding variables. In this study, we will conduct systematic and comprehensive relative survival and mean residual life analyses of HIV/AIDS patients with the most recent data from the United State National HIV/AIDS reporting system (HARS).

In chapter II, we describe the methods for data analysis. Chapter III includes data analysis and results. Chapter VI contains the discussion. Chapter V includes conclusions for this study. Chapters VI, VII, VIII, and IX are references, tables, figures, and appendixes respectively.

## CHAPTER II. METHODS

### 1. Data Source

#### 1.1. Definitions

**HIV-infected patients (HIV):** For most HIV-infected patients, **HIV infection time is unknown**. The HIV patient's diagnosis information in HARS is the date of first HIV positive test (HIV seroconversion time). The primary screening test of HIV

infections is called as Enzyme-Linked ImmunoSorbent Assay (ELISA). The confirming test of HIV infection is Western blot. After the positive ELISA test of HIV antigen/antibody, if the Western blot test is also positive, then the patient is considered HIV infected.

**AIDS patients (AIDS):** According to the CDC case definition, one has AIDS if he/she is infected with HIV and presents with one of the following:

- (1) A CD4<sup>+</sup> T-cell count below 200 cells/ $\mu$ l (or a CD4<sup>+</sup> T-cell percentage of total lymphocytes of less than 14%), or
- (2) He/she has one of the common defining opportunity illnesses [15].

## **1.2. Datasets Used in the Study**

Data of 1996-2003 diagnosed AIDS patients were from the U.S. national HARS, including all 50 United States; 1996-2003 diagnosed HIV cases were from 25 states (Alabama, Arizona, Arkansas, Colorado, Idaho, Indiana, Louisiana, Michigan, Minnesota, Mississippi, Missouri, Nevada, New Jersey, North Carolina, North Dakota, Ohio, Oklahoma, South Carolina, South Dakota, Tennessee, Utah, Virginia, West Virginia, Wisconsin, and Wyoming). Data used in this study satisfy the following criteria:

- **Inclusion criteria:**

- (1) AIDS patients were from the HARS.AS\_JUN06 dataset: N = 976,105 observations, 614 variables.
- (2) Laboratory data from the LAB.RA\_JUN06 dataset: N= 1,418,260 observations ,351 variables.
- (3) HIV patients from the HARS.HS\_JUN06 dataset: N= 266,437 observations, 613 variables.

- **Exclusion criteria:**

- (1) HIV/AIDS patients diagnosed at age <13 years.
- (2) HIV infection diagnosis year or month was missing.
- (3) AIDS diagnosis year or month was missing.
- (4) HIV/AIDS patients' death year or month was missing.
- (5) HIV/AIDS diagnosis year was <1996 or > 2003.

- **Follow up:** patients were followed up through the end of 2005. The maximum follow up time was 10 years.

The observed survival probabilities of AIDS patients were calculated by using the Kaplan-Meier method for the corresponding datasets. The expected survival probability was obtained by merging the national population data and mortality datasets sorted by age, sex, race/ethnicity, and calendar year. Both the population denominators and numerator mortality rate were from the National Center for Health Statistics [41]. The death rate in 2005 was not available and was therefore assigned to be the same value as that in 2004 in the analysis. The maximum age in the mortality dataset was 85 years. So if a person's age in the population dataset is over 85 years, his\her age will be assigned as 85 years.

## **2. Models and Estimation Approaches of Relative Survival and Mean Residual Life**

### **2.1. Models of Relative Survival**

For HIV/AIDS patient with covariate vector  $\mathbf{z}$ , the hazard rate at time  $t$  is modeled as the sum of the expected hazard rate  $\lambda^*(t; \mathbf{z})$  and the excess hazard rate  $v(t; \mathbf{z})$  due to a diagnosis of AIDS:

$$\lambda(t; \mathbf{z}) = \lambda^*(t; \mathbf{z}) + v(t; \mathbf{z}), \quad (1)$$

where the expected hazard (sometimes called the baseline hazard) is estimated from general-population mortality rates.

The corresponding survival function can be written as:

$$S(t; \mathbf{z}) = S^*(t; \mathbf{z}) r(t; \mathbf{z}), \text{ or } r(t; \mathbf{z}) = S(t; \mathbf{z}) / S^*(t; \mathbf{z}), \quad (2)$$

Where  $r(t; \mathbf{z})$  is the relative survival.

If the excess hazard rate,  $v(t; \mathbf{z})$ , is assumed to be a multiplicative function of the covariates, written as  $\exp(\mathbf{z}'\boldsymbol{\beta})$ , then the model for the relative survival of patients with HIV/AIDS diagnosis is :

$$\lambda(t; \mathbf{z}) = \lambda^*(t; \mathbf{z}) + \exp(\mathbf{z}'\boldsymbol{\beta}), \quad (3)$$

where  $\boldsymbol{\beta}$  is an unknown parametric vector. Based on the equation (3), estimation methods of relative survival were developed in references [42-44] with full likelihood and generalized linear model (GLM) approaches.

## 2.2. Estimation Approaches of Relative Survival Analysis [45]

Four approaches can be applied to estimate the relative survival with the dataset of AIDS to death (339,863 AIDS patients, diagnosed from 1996 to 2003, followed up through the end of 2005):

- (1) Grouped survival times, GLM with a binomial error structure ;
- (2) Grouped survival times, GLM with a Poisson error structure ;
- (3) Exact survival times, individual subject-band observations;
- (4) Exact survival times, collapsed data, GLM with a Poisson error structure.



The collapsed data used in the fourth approach contained the same number of observations as the grouped data sets. The distinction between collapsed data and grouped data is that, in the collapsed data,  $y$  (**person time, years at risk using exact times  $y_i$** ) and  $d^*$  ( **$d^*$  expected death number**) are based on the exact time at risk whereas these quantities are approximated for grouped data.

The selected models were fitted with AIDS patient data and the general-population mortality rates. The excess hazard ratios, or relative excess risks, are assumed to be an exponential function of unknown parameters [46]. An excess hazard ratio of, for example, 1.5 for a patient with HIV viral load 2 unit, compared to a patient with HIV viral load 1 unit means that the excess risk of death will increase 50% if the HIV RNA copy number in peripheral blood increases 1 unit.

### **2.3 Relative excess hazard ratio and AIDS free survival proportion**

Model (4) was used to estimate the main effect for each variable adjusted for all of the other variables in the model. The first level of each variable was set up as reference level. After obtaining the main effect of each level, the relative excess hazard ratio was the exponentiated rate of the effect for the corresponding level to the referenced level. AIDS free survival proportion was calculated using the Kaplan Meier method.

### **2.4 Modeling and Estimation of HIV/AIDS Mean Residual Life**

Mean residual life is a measurement of the remaining life expectancy of a subject at time  $t$ , which is the remaining survival time given the subject surviving up to  $t$ . Dr. Qin and Dr.

Zhao's methodology was employed to model the distribution of HIV/AIDS mean residual life [40].

### **3. Variable Selection and Model Checking**

The AIDS to death cohort dataset was used for model building. The dataset has 976105 observations, 614 variables, plus the variables (CD4<sup>+</sup> T cell count, HIV viral load as well as CD4<sup>+</sup> T-cell count detection times) from the laboratory dataset; the total variable count was 663. Based on published papers [22, 38], sex, age group, race, transmission category, HIV/AIDS diagnosis year, CD4<sup>+</sup> T-cell count, viral load were selected for inclusion in our model. These variables have a close relationship with disease development and clinical progression. In our dataset, the HIV viral load has a large amount of missing (57.3%); CD4<sup>+</sup> T-cell count information has 14.2% missing (see the discussion).

The selected approaches (see table 5 for detail) (1), (2) and (4) are generalized linear models. They can be checked by the deviance statistic which is a measure of goodness-of-fit [43]. Excess hazard in approaches (1), (2), and (4) require the assumption of proportional hazard in a pre-specified interval band, so each variable will be tested using the log rank test.

All the analyses were implemented using SAS 9.1 program, except that the analysis of HIV/AIDS means residual life was performed using S-plus version 7.0.

## **CHAPTER III. DATA ANALYSIS AND RESULTS**

Three datasets were generated: 339,863 AIDS cases were selected for the study of AIDS to death; for the HIV sero-conversion (HIV) to death study, 122,391 HIV-infected

patients were selected; for the HIV sero-conversion (HIV) to AIDS study, 122,379 HIV-infected patients were selected.

### **1. Variable selection and survival curve analysis**

The log rank test was used to check the significance of seven preliminary selected variables: race/ethnicity (non-Hispanic white (white), non-Hispanic black (black), Hispanic, Asian/Pacific Islander (A/PI), and American Indian/Alaska Native (AI/AN)), age at diagnosis, sex (male and female), diagnosis year, HIV transmission category (men who have sex with men(MSM), male drug user, female drug user, men sex men and drug user, female high risk heterosexual contact, male high risk heterosexual contact, CD4<sup>+</sup> T-cell count, and HIV viral load.

For variable selection, survival curves were plotted by each risk factor to check equality over strata for the proportional hazard assumption. Figure 2 shows that Blacks and AI/AN had worse survival probabilities compared with A/PI. Figure 3 shows that there is an obvious monotonous trend among the age groups. The survival probability of AIDS patients decreases when age increases. Figure 4 shows that the survival probabilities from highest to lowest transmission categories are: male to male sex (MMS, similar to MSM), female high-risk heterosexual contact (HCF), male high-risk heterosexual contact (HCM), unknown risk group, MSM and injection drug use (IDU), male injection drug use (IDUM), female injection drug use (IDUF). Injection drug users are the highest risk AIDS transmission group. Additionally, from Figure 4, the assumption of proportional hazard was not valid for the variable transmission category. After adding interaction terms (MSM with IDU, HCM vs HCF), the assumption of proportional hazard was valid. From Figure 5,

we observe that patient survival was better when AIDS was diagnosed later in time (e.g., 2003 versus 1996). Figure 6 shows that the survival curves of males and females are very similar. Figure 7 indicates that the survival probability of AIDS patients would improve with increased CD4<sup>+</sup> T-cell count (first CD4<sup>+</sup> T-cell count within 6 months of AIDS diagnosis). Figure 8 shows that with the increased viral load (first available record), the survival probability among AIDS diagnosed patients decreased.

## **2. Relative survival from AIDS diagnosis to death**

Table 1 provides the results of relative survival probabilities from AIDS diagnosis to death. There were 339,863 persons diagnosed with AIDS at age  $\geq 13$  years in the 50 states and the District of Columbia from 1996 to 2003. Forty eight percent of the AIDS patients were black, 31% were white, 20% were Hispanic, and less than 1% were A/PI or AI/AN (Table 1, columns 2, 3). The number of male AIDS patients was 3 times the number of females. Seventy two percent of patients were from 30 to 50 years old. Twenty seven percent were diagnosed with AIDS at very low CD4<sup>+</sup> T-cell counts ( $<50$  cells/uL).

Hispanic and Black AIDS patients have lower relative survival probability compared with white patients. AI/ANs have the lowest relative survival probability. Generally, with the increase of age at AIDS diagnosis and viral load, the relative survival rate among AIDS-diagnosed patients decreases; while with increasing CD4<sup>+</sup> T-cell count in peripheral blood, the relative survival probability increases.

The difference of the cumulative relative survival probability between men and women was small; women had slightly higher survival at one year and slightly worse at

three years after AIDS diagnosis compared with men. One and three years after AIDS diagnosis, MSM had the highest cumulative relative survival probability compared with all the other transmission categories. Injection drug users had the worst relative survival probability. From 1996 to 2003, each year the survival of newly diagnosed AIDS patients was slightly better, even though year to year improvements were not significant.

### **3. Relative survival from HIV infection diagnosis to death**

Table 2 gives the results from HIV diagnosis to death for cases diagnosed from 1996 to 2003 in 25 states using integrated HIV/AIDS surveillance reporting system. Fifty percent of HIV patients were black, 34% were white, 8.5% were Hispanic, and less than 1% was Asian/Pacific Islander or American Indian/Alaska Native (Table 2, columns 2, 3). The number of males was almost 3 times the number of female HIV patients (males account for 72%, females 28%). Over sixty percent of HIV patients were aged from 30-49 years; twenty three percent were 20-29 years old. Sixteen percent were diagnosed with HIV at very low CD4<sup>+</sup> T-cell count (<50 cells/uL).

The differences of the cumulative relative survival rates among race/ethnicity, sex and diagnosis year can be seen in Table 2 (column 4 for 1-year and column 6 for 3-year cumulative survival). Similar to AIDS, HIV patients diagnosed at older ages had lower relative survival rates compared with those diagnosed at younger ages. Among different transmission categories, the injection drug users had the lowest relative survival. HIV infected patients with higher CD4<sup>+</sup> T-cell count also had better survival, but this information was missing for 25% of patients. If the HIV patients had lower viral loads,

they would have higher relative survival. Fifty one percent of patients did not have viral load test results.

#### **4. Relative excess hazard ratio**

Table 3 presents the results obtained by using the estimation approach (4). Columns 2 and 4 present the relative excess hazard rate for AIDS and HIV patients respectively. The excess hazard rate of death for black, Hispanic, and AI/AN with AIDS and HIV patients were investigated after adjustment for sex, age, transmission category, diagnosis year, CD4<sup>+</sup> T-cell count as well as HIV viral load. For each category, the first level was used as the reference level. For race, HIV infected AI/AN patients have the highest relative excess hazard rate (1.41 and 1.37 for AIDS and HIV, respectively). A/PI AIDS cases had the lowest relative excess hazard rate (0.95 and 0.91 for AIDS and HIV, respectively). In contrast to the corresponding circumstance in relative survival analysis, the relative excess hazard rates of AIDS to death and HIV to death increased with the increase of age at AIDS diagnosis. If HIV viral load increased, then the relative hazard rate for AIDS and HIV patients increased; while with increasing CD4<sup>+</sup> T-cell count, the relative excess hazard rates decreased. From 1996 to 2003, each year the relative excess hazard rate of newly diagnosed AIDS and HIV decreased slightly. For certain variables (such as race/ethnicity), compared with HIV, all the AIDS patients had higher relative excess hazard rate except age group. The relative excess hazard rate for the oldest 60+ year AIDS patients was 6.93 (95% CI: 5.94 to 8.09); while for HIV patients, it was 20.14 (95% CI: 14.5 to 28.1).

## **5. AIDS free survival proportion from HIV to AIDS diagnosis**

From Table 4, we can see that among 122,379 patients diagnosed with HIV from 1996 to 2003, 64,423 progressed to AIDS before the end of 2003. Among 2003 HIV diagnoses, after 12 months, 60% of the patients had not progressed to AIDS; after 36 months, 56% had not developed AIDS. Sixty percent of women had not progressed to AIDS 3 year after HIV diagnosis; while 50% HIV infected males had developed AIDS during that time frame.

HIV infected A/PIs were least likely to have developed AIDS 3 years after HIV diagnosis compared with Blacks, Hispanics, or American Indians/Alaska Natives. The AIDS-free rate of patients diagnosed with HIV at younger ages was higher compared with those diagnosed with HIV infection at older ages for both 1 and 3 year after HIV diagnosis. Among transmission categories that affect both men and women, the AIDS-free survival rate at 36 months was higher for women than for men. The AIDS-free survival rate did not change substantially between diagnosis year, but did improve from 1996 to 2003 (50.0 to 55.9 for 3 years after HIV diagnosis). AIDS-free survival probability decreased with increasing HIV viral load, when HIV viral load was up to 5, the AIDS-free survival probability decreased dramatically.

## 6. Model comparison and model checking

Table 5 shows that the main effect for each variable had no obvious difference when estimated by different models (the values of main effect are similar for a single variable from different models). Model 1 was developed based on equation 9. Assuming that the number of deaths in each interval follows a binomial distribution, the outcome is  $l_{ki} - d_{ki}$  (the number of patients surviving the interval), with binomial error structure with denominator  $l_{ki}$ . Model 2 was built based on grouped events (deaths) with approximate survival time. The number of deaths in a pre-specified time interval band (length = 1 year) follows a Poisson distribution with approximate survival time. Model 3 was developed based on the assumption that the number of deaths in a pre-specified interval band (length = 1 year) follows a Poisson distribution, but the data are individual level and survival time is the exact time. Model 4 was generated from the assumption that the number of deaths in a pre-specified interval band (length = 1 year) follows a Poisson distribution, the data are from grouped level and survival time is exact. The goodness of fit test for selected approaches has p a value greater than 0.05, so there is not enough evidence to reject the current models, thus our data fit the models well.



## CHAPTER 4. DISCUSSION

HIV infection prevention is one of the most challenging tasks for public health workers. In order to improve the quality of life of HIV-infected patients and extend the interval from HIV infection /AIDS diagnosis to death, the US Department of Health and Human Services has implemented the use of many prophylaxis, intervention strategies and methodologies, such as mass media campaigns, peer education about HIV transmission, treatment of other sex transmitted infection, condom social marketing, safe blood provision, and needle exchange. Survival analysis is not only a conventional technology but also a frequently implemented technique to evaluate the efficiency of new treatment and prevention methods. Relative survival was a very popular cancer survival analysis technique; the first advantage is that it does not need the information of the specific cause of death. Another advantage is that relative survival analysis excludes the effect of baseline mortality (such as old age) on the survival rate under study [47, 48]. Compared with cancer studies, this population-based survival analysis recently began to attract the attention of HIV/AIDS researchers [37, 38]. HIV/AIDS relative survival analysis will provide a measure of the excess mortality experienced by patients diagnosed with HIV infection or AIDS, irrespective of whether or not the excess mortality is directly or indirectly attributable to HIV infection.

Our results showed that the relative excess hazard of death was higher for blacks, American Indian /Alaska Natives, as well as Hispanics with HIV infection and AIDS diagnosis as compared with whites. Excess risk for death also was higher for persons with a diagnosis of HIV or AIDS at an older age, and lower CD4<sup>+</sup> T-cell count at diagnosis.

Additionally, if HIV/AIDS patients have a higher viral load, they will have a higher relative excess hazard rate compared to the HIV/ AIDS patients, who have a lower HIV viral load.

CD4<sup>+</sup> T-cell count is one of the most commonly used surrogates for disease progression [49, 50] and disease severity/immune dysfunction as well as a criterion for evaluation of the effect of clinical and prophylaxis intervention applied in HIV/AIDS patients. Usually, the variable data on this marker is characterized by unequal numbers of unevenly spaced repeated observations. In our data set, a patient can maximally have up to 20 times CD4<sup>+</sup> T-cell count values. Almost half of patients have no CD4<sup>+</sup> T-cell count information (48%) (the time interval between any two time CD4<sup>+</sup> T-cell count values is different for each patient). CD4<sup>+</sup> T-cell count has many components of variability: measurement error, short term variability, and long term trend. To avoid the correlation among repeated CD4<sup>+</sup> T-cell count in one subject, we used the first CD4<sup>+</sup> T-cell count in our model after disease diagnosis but within 6 months of HIV infection or AIDS diagnosis. Further investigation is needed to determine the effect of the missing data of CD4<sup>+</sup> T-cell count on results. Since 1996, using HIV viral load as an independent predictor variable to predict outcome (disease progression, treatment) has attracted wide attention [1, 51]. The results from our study are similar to previous publications: the higher HIV RNA viral load, the lower survival probability [1, 51].

Similar results were obtained from our analysis and other research about the effect of diagnosis age for HIV/AIDS as an important determinant of disease progression. Our results show that younger patients have the advantage of surviving longer than older patients. It was reported that younger age favored CD4<sup>+</sup> T-cell restoration upon HAART,

which is consistent with younger persons having good thymic function for CD4<sup>+</sup> T-cell generation [1]. With respect to the transmission categories, IDUs had lower relative survival probability (higher relative hazard rate) 36 months after HIV infection and AIDS diagnosis than other transmission categories. Independent of HIV infection, IDU patients have an increased risk of death from overdose, violent causes, HCV co-infection, liver toxicity, or a decline of adherence of treatment [38]. All of these risk factors might contribute to this transmission category pattern. Between men and women, there was no significant difference in relative excess risk for death and the relative survival.

The disparities of the AIDS-free survival probability in our analysis may relate to access or adherence to treatment or diagnosis delay, or lack of adequate care. If AIDS diagnosis is delayed in disadvantaged population groups (e.g., Hispanics), then it is difficult to obtain a precise estimate of the AIDS-free probability difference (between Hispanics, blacks, and whites) if it exists.

The advantage of relative survival analysis is obvious for HIV/AIDS survival studies. In the cohort dataset of HIV to death, it was found that 15% of HIV patients were dead before AIDS diagnosis. These patients were not dead from AIDS defining events; secondly, we do not need the cause-specific information of death, which can be biased toward under ascertainment of HIV/AIDS as a cause of death, to perform relative survival analysis.

For model selection, generalizability depends on independent predictors. Our model included the most important progress factors so far identified. From the point of model simplicity, we did not include genetic factors (MIDR1, 3435C/T polymorphism, CCR5)32 mutation), which have been shown to affect response to antiretroviral therapy [52]. Consideration of the specificity of survival data (events may count at pre-specified

intervals or at the exact survival time), four approaches were tested to model the relative excess hazard of death by using the AIDS to death cohort dataset. Similar results for estimating the relative survival were obtained when using different models. This is not surprising since similar methods and datasets were used to model the relative survival of AIDS patients. Model (1) was based on a binomial assumption. Model (2) used grouped data (events were counted in approximate survival time). Model (3) used individual level of data. Finally, model (4) used collapsed events obtained from exact survival time.

The relative survival results of Table 1, 2 and 3 were obtained with the generalized linear model (model 4) based on collapsed data with exact survival times and a Poisson assumption. The resulting Goodness of fit test showed that our data fit well for the selected model 4, with a deviance/df of 0.7799. Meanwhile, two term interactions were tested. It was found that there were significant interactions between age and year of follow-up (fu), age and diagnosis year (dxyr). The main effects of follow-up year and sex are not significant in the model, thus, they should be excluded from the final model. The final model may include age, race/ethnicity, transmission category, CD4<sup>+</sup> T-cell count, viral load, and the two interactions fuXage and fuXdxyr.

The approach developed by Qin and Zhao [40] was used to estimate the mean residual life for HIV and AIDS patients after survival 12 months after diagnosis. Our results showed that the estimated mean residual life for HIV and AIDS patients followed the same pattern as the survival probability: if the HIV and AIDS patient groups had a higher relative survival probability (or survival probability), then they had a longer mean residual life (white patients); if the patients had a lower survival probability (injection drug users), then they had a shorter mean residual life. Compared to the results of survival probability,

the mean residual life was shorter than expected, especially for newly diagnosed HIV and AIDS patients. This bias may be associated with the large amount of censoring and the dependency between censoring and survival time. More than 80% of AIDS cases diagnosed in 2003 were censored. A case with a longer survival time has a higher probability of being censored. Therefore, the results on mean residual life were not shown. We will continue our efforts to solve this problem in estimating the mean residual life.

Survival analysis is a widely accepted method to evaluate the efficacy of new treatment and prophylaxis strategies. In our study, the nation-wide HIV/AIDS data was used to estimate the mean residual life. Mean residual life has the potential to be easily understood by clinicians. For example, at the time point ( $t$ ) to implement a new treatment or a new HIV vaccine (after HIV/AIDS patients surviving 12 months), the mean residual life is the improved life time of HIV and AIDS patients after using the new treatment or the new vaccine.

Since only the first time detection data of  $CD4^+$  T-cell count and viral load was included in the model (both  $CD4^+$  T-cell count and viral load can be used to reflect disease stage or severity). One of the limitations in our study was that we did not have information on the severity of the HIV infection caused disease. Furthermore, over 50% of the viral load information was unknown. Secondly, for most patients, HIV infection time was unknown, we used HIV seroconversion time to replace HIV infection time, which may cause a slight difference in the survival analysis from HIV infection to AIDS/death. Finally, The data used to assess relative survival among persons with an HIV diagnosis (with or without a concurrent diagnosis with AIDS) or assess disease progression from HIV

diagnosis to AIDS included 25 states; and it is not known whether results can be extrapolated to the United States as a whole.

Further efforts are needed to address whether differences in survival and mean residual life are the result of treatment and/or prevention interventions.  $CD4^+$  T-cell count obtained at different time points need to be incorporated in models describing the full scenario of disease progression. HIV viral load as an independent variable to predict survival and mean residual life deserves further investigation.

## **Chapter 5. CONCLUSIONS**

- Generalized linear model with Poisson error structure was an appropriate approach for estimating the relative survival of HIV/AIDS patients.
- Younger persons, MSM, and whites have better survival compared to older, IDUs, blacks, and Hispanics after HIV or AIDS diagnosis.
- HIV diagnosed patients have higher relative excess hazard compared with AIDS diagnosed patients.
- When viral load was larger than  $\log_5$ , the AIDS-free survival probability 36 months after HIV diagnosis decreased considerably.
- HIV viral load is a very important risk factor affecting HIV/AIDS survival (its value of the estimated main effect and the relative excess hazard is highest).
- It is worthwhile to further the analysis of mean residual life for HIV/AIDS patients.

## REFERENCES

1. David M Knipe, Peter M Howley, Martin MA, et al, Fields Virology, 4<sup>th</sup> edition, Lippincott – Raven, Philadelphia , USA 2001.
2. Anonymous. Pneumocystis pneumonia: Los Angeles. MMWR Morb Mortal Wkly Rep 1981;30:250–252.
3. Gottlieb MS, Schroff R, Schanker HM, et al. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: Evidence of a new acquired cellular immunodeficiency. N Engl J Med 1981;305:1425–1431.
4. Essex M, McLane MF, Lee TH, et al. Antibodies to cell membrane antigens associated with human T-cell leukemia virus in patients with AIDS. Science 1983;220:859–862.
5. Barr-Sinoussi F, Chermann JC, Rey F, et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science, 1983, 220:868–871.
6. Levy DN, Refaeli Y, MacGregor RR, et al. Serum Vpr regulates productive infection and latency of human immunodeficiency virus type 1. Proc Natl Acad Sci U S A, 1994;91:10873–10877.
7. Coffin J, Haase A, Levy JA, et al. Human immunodeficiency viruses [letter]. Science 1986;232:697.
8. Clavel F, Gutard D, Brun-Vzinet F, et al. Isolation of a new human retrovirus from West African patients with AIDS. Science 1986;233:343–346.
9. UNAIDS Joint United Nations Programme on HIV/AIDS. AIDS epidemic update: December. Geneva: UNAIDS, 2000.
10. Glynn MK, Rhodes P. Estimated HIV prevalence in the United States at the end of 2003 [Abstract T1-B1101]. Proceedings of the 2005 National HIV Prevention Conference; 2005 Jun 12-15; Atlanta,GA.
11. Vittinghoff E, Douglas J, Judson F, et al. Per-contact risk of human immunodeficiency virus transmission between male sexual partners. Am J Epidemiol 1999;150:306–311.
12. Palella FJ Jr, Delaney KM, Moorman AC, et al. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. N Engl J Med 1998;338:853–860.
13. The White House Office of National AIDS Policy. Report on the presidential mission

on children orphaned by AIDS in sub-Saharan Africa: Findings and plan of action. Washington, DC: The White House Office of National AIDS Policy, 1999.

14. Kinloch-de Loes S, de Saussure P, Saurat JH, et al. Symptomatic primary infection due to human immunodeficiency virus type 1: Review of 31 cases. *Clin Infect Dis* 1993;17:59–65.
15. Centers for Disease Control and Prevention. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *MMWR Morb Mortal Wkly Rep* 1992;41:1–19.
16. Piatak M, Saag MS, Yang LC, et al. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science* 1993;259:1749–1754.
17. Anonymous. Interpretation and use of the Western blot assay for serodiagnosis of human immunodeficiency virus type 1 infections. *MMWR Morb Mortal Wkly Rep* 1989;38:1–7.
18. Koot M, Keet IPM, Vos AHV, et al. Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. *Ann Intern Med* 1993;118:681–688.
19. Masur H, Ognibene F, Yarchoan R, et al. CD4 counts as predictors of opportunistic pneumonias in human immunodeficiency virus (HIV) infection. *Ann Intern Med* 1989;111:223–231.
20. Giorgi JV, Liu Z, Hultin LE, et al. Elevated levels of CD38–CD8+ T cells in HIV infection add to the prognostic value of low CD4+ T cell levels: Results of 6 years of follow-up. *J Acquir Immune Defic Syndr* 1993;6:904–912.
21. Carpenter CC, Fischl MA, Hammer SM, et al. Antiretroviral therapy for HIV infection in 1998: Updated recommendations of the International AIDS Society-USA Panel. *JAMA* 1998;280:78–86.
22. Department of Health and Human Services/Henry J. Kaiser Family Foundation Panel on Clinical Practices for Treatment of HIV Infection. Guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents: Current guidelines available at [www.hivatis.org](http://www.hivatis.org). *Ann Intern Med* 1998;128:1079–1100.
23. Cameron DW, Heath-Chiozzi M, Danner S, et al. Randomised placebo-controlled trial of zidovudine in advanced HIV-1 disease: The Advanced HIV Disease Zidovudine Study Group. *Lancet* 1998;351:543–549.
24. Finzi D, Blankson J, Siliciano JD, et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med* 1999;5:512–517.



25. Putter H, Heisterkamp SH, Lange JM, A Bayesian approach to parameter estimation in HIV dynamic models. *Statistics in Medicine* 2002, 21 (15):2199-2214.
26. Follmann D. Augmented designs to assess immune response in vaccine trials. *Biometrics*. 2006 Dec;62(4):1161-9.
27. Cole SR, Hernán MA, Margolick JB, Cohen MH, Robins JM. Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *Am J Epidemiol*. 2005 Sep 1;162(5):471-8. Epub 2005 Aug 2.
28. Glynn MK, Rhodes P. Estimated HIV prevalence in the United States at the end of 2003 [Abstract]. Presented at the National HIV Prevention Conference, June 12-15, 2005; Atlanta, Georgia.
29. Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *JAMA* 1998; 280:42–48.
30. Karon, K.M., Song, R., Kaplan, E., and Brookmeyer, R. Estimating HIV incidence from HIV/AIDS surveillance data and less-sensitive HIV test results. Submitted for publication.
31. Hall HI, Song R, Rhodes P, Prejean JG, An Q, Lee LM, Karon J, Brookmeyer R, Kaplan EH, McKenna MT, Janssen RS, and the HIV Incidence Surveillance Group. HIV incidence in the United States, 2005. Submitted for publication.
32. Posner SJ, Myers L, Hassig SE, et al, Estimating HIV incidence and detection rates from surveillance data, *epidemiology*, 2004, 15(2):164-72.
33. Mittlbock M, Whitehead J. The interpretation of clinical trials of immediate versus delayed therapy. *Lifetime Data Anal.* 1998;4(3):253-63.
34. Betensky RA, Finkelstein DM. A non-parametric maximum likelihood estimator for bivariate interval censored data. *Stat Med.* 1999 Nov 30;18(22):3089-100.
35. Hogan JW, Laird NM. Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Stat Methods Med Res.* 1998 Mar;7(1):28-48.
36. Ederer F, Axtell J., The relative survival rate: a statistical methodology. *Natl Cancer Inst Monogr.* 1961 Sep;6:101-21.
37. Fraser S, Clarke K. Workshop report. National initiative to improve cancer survival information. *Chronic Dis Can.* 1998;19(2):62-70.
38. Hall HI, McDavid K, Ling Q, Sloggett A. *Related Articles*,\_Determinants of

- progression to AIDS or death after HIV diagnosis, United States, 1996 to 2001. *Ann Epidemiol.* 2006 Nov;16(11):824-33.
39. Lohse N, Hansen AB, Pedersen G, Kronborg G, Gerstoft J, Sorensen HT, Vaeth M, Survival of persons with and without HIV infection in Denmark, 1995-2005, *Ann Intern Med.* 2007 Jan 16;146(2):87-95.
  40. Qin, G., Zhao, Y., Empirical likelihood inference for the mean residual life under random censorship. *Statistics & Probability Letters* (2006), doi:10.1016/j.spl.2006.09.018
  41. National Center for Health Statistics. Bridged-Race Intercensal Estimates of the July 1, 1990-July 1, 1999, US Resident Population by County, Single-Year of Age, Sex, Race, and Hispanic Origin, Prepared by the US Census Bureau with support from the National Cancer Institute. Available :<http://www.cdc.gov/nchs/about/major/dvs/popbridge/popbridge.htm>.
  42. Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* 1990; 9:529 –538.
  43. Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987; 36:309 –317.
  44. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd ed). Chapman & Hall: London, 1989.
  45. Paul W. Dickman, Andy Sloggett, Michael Hills and Timo Hakulinen, Regression models for relative survival, *Statist. Med.* 2004; 23:51–64.
  46. Suissa S. Relative excess risk: an alternative measure of comparative risk. *American Journal of Epidemiology* 1999; 150:279–282.
  47. Skiest DJ, Rubinstien E, Carley N, Gioiella L, Lyons R. The importance of comorbidity in HIV-infected patients over 55: A retrospective case-control study. *Am J Med.* 1996;101:605–611.
  48. Manfredi R. HIV disease and advanced age: An increasing therapeutic challenge. *Drugs Aging.* 2002;19:647–669.
  49. CASCADE Collaboration. Short-term risk of AIDS according to current CD4 cell count and viral load in antiretroviral drug-naïve individuals and those treated in the monotherapy era. *AIDS.* 2004;18:51–58.
  50. Phillips AN, Lee CA, Elford J, Webster A, Janossy G, Timms A, et al. More rapid progression to AIDS in older HIV-infected people: the role of CD4 T-cell counts. *J*

AIDS. 1991;4:970–975.

51. John W. Mellors, Charles R. Rinaldo Jr., et al, Prognosis in HIV infection predicted by the quantity of virus in plasma. *Science*, 1996, 272: 1167-1170.
52. Fellay J, Marzolini C, Medden ER, et al, Response to antiretroviral treatment in HIV1 infected individuals with allelic variants of the multidrug resistance transporter, *Lancet*, 2002, 359:30-36.

**Table 1 Relative survival after AIDS diagnosed from 1996 to 2003,  
by 50 US States**

Characteristics	Frequency	Percent	CR1	CICR1	CR3	CICR3
Race/ethnicity						
non-Hispanic white	104104	30.6	88.8	(88.5, 88.9)	81.0	(80.7, 81.3)
non-Hispanic black	163154	48.0	86.6	(86.4, 86.7)	75.2	(75.0, 75.5)
Hispanic	68196	20.1	86.4	(86.2, 86.7)	78.6	(78.3, 79.0)
Asian/Pacific Islander	3063	0.9	88.0	(86.8, 89.1)	82.8	(81.3, 84.2)
American Indian/Alaska Native	1346	0.4	85.4	(83.3, 87.1)	73.2	(70.3, 75.9)
Age at diagnosis (years)						
13-19	2728	0.8	96.2	(95.4, 96.8)	89.6	(87.6, 91.3)
20-29	43386	12.8	92.2	(92.0, 92.5)	83.8	(83.3, 84.2)
30-39	137193	40.4	90.0	(89.8, 90.1)	81.2	(80.9, 81.4)
40-49	107259	31.6	85.8	(85.7, 86.1)	76.0	(75.7, 76.3)
50-59	36574	10.8	80.2	(79.8, 80.7)	70.0	(69.4, 70.4)
60+	12723	3.7	69.2	(68.3, 70.0)	59.2	(58.2, 60.1)
Transmission category						
MSM	120762	35.5	89.8	(89.6, 90.0)	82.8	(82.5, 83.0)
IDU Male	50866	15.0	83.4	(83.0, 83.7)	70.4	(70.0, 70.9)
IDU Female	23648	7.0	86.2	(85.8, 86.6)	71.0	(70.3, 71.7)
MSM+IDU	18010	5.3	90.8	(90.3, 91.2)	79.8	(79.0, 80.4)
HC Male	40054	11.8	81.4	(80.9, 81.7)	73.2	(72.8, 73.8)
HC Female	20359	6.0	85.2	(84.7, 85.7)	76.0	(75.3, 76.6)
Unknown	66164	19.5	89.0	(88.8, 89.3)	79.6	(79.2, 79.9)
Sex						
Male	255371	75.1	87.0	(86.9, 87.2)	78.2	(78.0, 78.4)
Female	84492	24.9	87.8	(87.5, 87.9)	76.4	(76.1, 76.8)
Diagnosis year						
1996	58286	17.2	84.2	(84.0, 84.6)	71.8	(71.4, 72.2)
1997	47637	14.0	86.8	(86.4, 87.0)	75.4	(75.0, 75.9)
1998	40932	12.0	86.6	(86.2, 86.9)	76.2	(75.8, 76.7)

1999	39067	11.5	87.2	(86.8, 87.5)	77.4	(76.9, 77.9)
2000	39392	11.6	87.8	(87.4, 88.1)	78.4	(77.9, 78.8)
2001	38210	11.2	88.2	(87.9, 88.6)	79.6	(79.2, 80.1)
2002	38073	11.2	88.8	(88.5, 89.1)	81.4	(81.0, 81.9)
2003	38266	11.3	90.0	(89.6, 90.3)	86.0	(85.5, 86.3)
First CD4 CNT						
<50	91730	27.0	81.0	(80.8, 81.3)	69.0	(68.7, 69.4)
50-99	43815	12.9	88.2	(87.9, 88.5)	77.8	(77.4, 78.3)
100-199	96745	28.5	93.6	(93.4, 93.8)	85.4	(85.1, 85.7)
>=200	59497	17.5	93.8	(93.5, 93.9)	86.8	(86.4, 87.1)
Unknown	48076	14.2	77.2	(76.8, 77.6)	67.8	(67.3, 68.2)
Log10(viraload)						
1-2	14052	4.1	98.8	(98.6, 99.1)	96.8	(96.4, 97.3)
3	23872	7.0	97.6	(97.3, 97.8)	94.0	(93.6, 94.4)
4	30966	9.1	96.4	(96.2, 96.6)	91.0	(90.6, 91.4)
5	53801	15.8	93.8	(93.7, 94.1)	86.0	(85.6, 86.3)
6-8	22321	6.6	89.8	(89.5, 90.3)	79.8	(79.1, 80.4)
Unknown	194851	57.3	81.4	(81.3, 81.7)	69.8	(69.6, 70.1)
Total	339863	100.00	87.2	(87.1, 87.3)	77.8	(77.6, 77.9)

CR1: 1-year cumulative relative survival proportion.

CICR1: 95% confidence interval of 1 –year cumulative survival rate.

CR3: 3-year cumulative relative survival proportion.

CICR3: 95% confidence interval of 3 –year cumulative survival rate.

MSM: men who have sex with men.

MSM+IDU: men who have sex with men and injection drug user;

HC: heterosexual contact with a person at high-risk for or a person with HIV infection or AIDS;

AIDS: acquired immunodeficiency syndrome; HIV: human immunodeficiency virus.

First CD4 CNT: First CD4<sup>+</sup> T cell count within 6 months of AIDS diagnosis.

Viraload: HIV RNA copy number.

**Table 2 Relative survival after HIV diagnosis from 1996 to 2003,  
by 25 US States**

Characteristics	Frequency	Percent	CR1	CICR1	CR3	CICR3
Race/ethnicity						
non-Hispanic white	42120	34.4	93.4	(93.1, 93.6)	89.2	(88.9, 89.6)
Non-Hispanic Black	68386	55.9	93.4	(93.1, 93.5)	87.2	(87.0, 87.6)
Hispanic	10358	8.5	93.2	(92.7, 93.7)	89.2	(88.4, 89.8)
Asian/Pacific Islander	636	0.5	93.6	(91.5, 95.3)	91.6	(89.1, 93.7)
American Indian/Alaska Native	891	0.7	92.2	(90.1, 93.7)	85.6	(82.6, 88.1)
Age at diagnosis( years)						
13-19	3822	3.1	99.4	(99.0, 99.5)	98.2	(97.3, 98.9)
20-29	27824	22.7	97.8	(97.7, 98.0)	95.2	(94.9, 95.5)
30-39	46412	37.9	94.6	(94.4, 94.9)	90.2	(89.8, 90.5)
40-49	31021	25.4	91.0	(90.6, 91.3)	84.6	(84.1, 85.0)
50-59	9949	8.1	85.8	(85.0, 86.4)	77.8	(76.9, 78.7)
60+	3363	2.8	73.4	(71.9, 75.0)	64.6	(62.6, 66.4)
Transmission category						
MSM	43745	35.7	94.6	(94.4, 94.9)	90.6	(90.3, 91.0)
IDU Male	10859	8.9	89.8	(89.2, 90.4)	80.6	(79.6, 81.4)
IDU Female	5276	4.3	92.8	(92.1, 93.6)	84.4	(83.1, 85.5)
MSM+IDU	4248	3.5	95.0	(94.3, 95.7)	88.6	(87.3, 89.6)
HC Male	8756	7.2	94.6	(94.1, 95.1)	89.2	(88.5, 90.1)
HC Female	16315	13.3	96.4	(96.0, 96.6)	92.0	(91.5, 92.5)
Unknown	33192	27.1	90.6	(90.4, 91.0)	85.6	(85.2, 86.1)
Sex						
Male	88578	72.4	92.8	(92.6, 93.0)	87.6	(87.4, 87.9)
Female	33813	27.6	94.6	(94.4, 94.9)	89.4	(89.1, 89.8)
Diagnosis year						
1996	17716	14.5	92.0	(91.6, 92.4)	84.8	(84.1, 85.4)
1997	16569	13.5	93.2	(92.9, 93.6)	86.8	(86.2, 87.4)

1998	15229	12.4	92.8	(92.4, 93.2)	87.2	(86.5, 87.7)
1999	14324	11.7	93.2	(92.7, 93.5)	87.6	(86.9, 88.2)
2000	14924	12.2	93.6	(93.3, 94.1)	88.6	(88.0, 89.2)
2001	14543	11.9	93.6	(93.2, 94.0)	88.6	(88.0, 89.2)
2002	14862	12.1	94.0	(93.6, 94.4)	90.4	(89.8, 91.0)
2003	14224	11.6	94.2	(93.7, 94.5)	92.4	(91.9, 92.9)
First CD4 CNT						
<50	20443	16.7	83.0	(82.4, 83.4)	71.8	(71.0, 72.4)
50-99	8628	7.1	90.0	(89.3, 90.6)	82.0	(81.0, 82.9)
100-199	15289	12.5	94.8	(94.4, 95.2)	88.0	(87.4, 88.7)
>=200	45366	37.1	98.8	(98.6, 98.8)	96.4	(96.1, 96.5)
Unknown	32665	26.7	92.4	(92.1, 92.7)	88.8	(88.4, 89.2)
Log10(viraload)						
1-2	3571	2.9	99.4	(99.0, 99.7)	97.8	(97.1, 98.5)
3	10754	8.8	99.2	(98.9, 99.3)	97.2	(96.7, 97.6)
4	16299	13.3	98.8	(98.5, 98.9)	96.4	(96.0, 96.8)
5	21238	17.4	96.0	(95.8, 96.4)	90.8	(90.3, 91.2)
6-8	7840	6.4	90.6	(89.8, 91.1)	82.4	(81.3, 83.3)
Unknown	62689	51.2	90.0	(89.7, 90.2)	83.8	(83.4, 84.1)
Total	122391	100.0	93.4	(93.2, 93.4)	88.2	(87.9, 88.3)

CR1: 1-year cumulative relative survival proportion.

CICR1: 95% confidence interval of 1 –year cumulative survival rate.

CR3: 3-year cumulative relative survival proportion.

CICR3: 95% confidence interval of 3 –year cumulative survival rate.

MSM: men who have sex with men.

MSM+IDU: men who have sex with men and injection drug user;

HC: heterosexual contact with a person at high-risk for or a person with HIV infection or AIDS;

AIDS: acquired immunodeficiency syndrome; HIV: human immunodeficiency virus.

First CD4 CNT: First CD4<sup>+</sup> T cell count within 6 months of AIDS diagnosis.

Viraload: HIV RNA copy number.

**Tables 3 Relative excess hazard rate of death within 3 years among patients with HIV or AIDS diagnosed from 1996 to 2003**

Characteristics	AIDS patients		HIV infection patients*	
	RR**	AIDS RR 95% CI	RR**	HIV RR 95% CI
<b>Race/ethnicity</b>				
non-Hispanic white	1.00		1.00	
non-Hispanic black	1.15	(1.13, 1.17)	1.02	(0.98, 1.07)
Hispanic	1.18	(1.15, 1.21)	1.07	(1.00, 1.15)
Asian/Pacific Islander	0.95	(0.86, 1.04)	0.91	(0.69, 1.21)
American Indian/Alaska Native	1.41	(1.26, 1.59)	1.37	(1.12, 1.67)
<b>Age at diagnosis (years)</b>				
13-19	1.00		1.00	
20-29	2.00	(1.71, 2.33)	2.55	(1.83, 3.55)
30-39	2.43	(2.09, 2.83)	4.81	(3.47, 6.68)
40-49	3.26	(2.80, 3.80)	7.19	(5.19, 9.98)
50-59	4.53	(3.89, 5.28)	11.17	(8.04, 15.5)
60+	6.93	(5.94, 8.09)	20.14	(14.5, 28.1)
<b>Transmission category</b>				
MSM	1.00		1.00	
IDU Male	1.51	(1.48, 1.55)	1.55	(1.46, 1.65)
IDU Female	1.65	(1.58, 1.74)	1.73	(1.55, 1.94)
MSM+IDU	1.08	(1.04, 1.13)	1.14	(1.02, 1.27)
HC Male	1.01	(0.98, 1.05)	0.83	(0.76, 0.90)
HC Female	1.15	(1.10, 1.20)	0.97	(0.88, 1.07)
Unknown	1.54	(1.50, 1.58)	1.37	(1.30, 1.44)
<b>Sex</b>				
Male	1.00		1.00	
Female	0.98	(0.95, 1.02)	0.92	(0.86, 0.98)
<b>Diagnosis year</b>				
1996	1.00		1.00	
1997	0.86	(0.84, 0.89)	0.89	(0.83, 0.95)



1998	0.86	(0.84, 0.88)	0.87	(0.82, 0.93)
1999	0.84	(0.81, 0.86)	0.84	(0.78, 0.90)
2000	0.86	(0.83, 0.88)	0.79	(0.74, 0.85)
2001	0.83	(0.81, 0.86)	0.82	(0.76, 0.88)
2002	0.77	(0.75, 0.80)	0.69	(0.64, 0.74)
2003	0.69	(0.66, 0.71)	0.66	(0.61, 0.71)
First CD4 CNT				
<50	1.00		1.00	
50-99	0.64	(0.62, 0.65)	0.57	(0.54, 0.61)
100-199	0.38	(0.38, 0.39)	0.36	(0.34, 0.39)
>=200	0.40	(0.39, 0.41)	0.13	(0.12, 0.14)
Unknown	0.92	(0.90, 0.94)	0.35	(0.34, 0.37)
Log10(viraload)				
1-2	1.00		1.00	
3	2.24	(1.92, 2.62)	1.79	(1.20, 2.69)
4	3.83	(3.31, 4.44)	2.80	(1.90, 4.12)
5	6.01	(5.21, 6.93)	5.11	(3.51, 7.43)
6-8	8.49	(7.35, 9.80)	7.84	(5.39, 11.4)
Unknown	13.87	(12.1, 16.0)	10.49	(7.24, 15.2)

\*HIV infection diagnosed with or without a concurrent diagnosis of AIDS

\*\*Relative excess hazard rate adjusted for all other variables (follow up period result not shown)

MSM: men who have sex with men; IDU: injection drug user;

MSMIDU: men who have sex with men and IDU: injection drug user;

HC: heterosexual contact with a person at high-risk for or a person with HIV infection or AIDS;

AIDS: acquired immunodeficiency syndrome; HIV: human immunodeficiency virus.

First CD4 CNT: First CD4<sup>+</sup> T cell count within 6 months of AIDS diagnosis.

Viraload: HIV RNA copy number.

**Table 4 AIDS free survival proportion from HIV to AIDS diagnosed from 1996 to 2003, by 25 US States**

Characteristics	Frequency	Percent	P1	CIP1	P3	CIP3
Race/ethnicity						
Non-Hispanic white	42120	34.4	58.3	(57.8, 58.7)	51.6	(51.1, 52.1)
Non-Hispanic black	68373	55.9	62.6	(62.3, 63.0)	53.29	(52.9, 53.7)
Hispanic	10360	8.5	55.0	(54.0, 56.0)	47.5	(46.5, 48.5)
Asian/Pacific Islander	636	0.5	58.3	(54.4, 62.1)	51.8	(47.6, 55.9)
American Indian/Alaska Native	890	0.7	60.9	(57.6, 64.0)	51.6	(48.0, 55.1)
Age at diagnosis (years)						
13-19	3822	3.1	86.7	(85.5, 87.7)	77.0	(75.4, 78.5)
20-29	27822	22.7	75.6	(75.1, 76.1)	66.8	(66.1, 67.4)
30-39	46397	37.9	59.2	(58.8, 59.7)	51.0	(50.4, 51.4)
40-49	31023	25.4	52.3	(51.8, 52.9)	44.2	(43.6, 44.8)
50-59	9947	8.1	46.2	(45.2, 47.1)	39.1	(38.1, 40.2)
60+	3368	2.8	39.2	(37.6, 40.9)	33.9	(32.2, 35.6)
Transmission category						
MSM	43744	35.7	57.7	(57.2, 58.1)	49.8	(49.3, 50.3)
IDU Male	10856	8.9	52.4	(51.5, 53.3)	42.8	(41.8, 43.8)
IDU Female	5273	4.3	65.9	(64.6, 67.2)	55.0	(53.4, 56.4)
MSM IDU	4245	3.47	60.9	(59.4, 62.3)	50.0	(48.2, 51.5)
HC Male	20971	17.1	59.0	(58.3, 59.7)	52.3	(51.6, 53.0)
HC Female	12222	10.0	68.3	(67.4, 69.1)	61.1	(60.2, 62.0)
Unknown	25068	20.5	64.9	(64.3, 65.4)	55.8	(55.1, 56.5)
Sex						
Male	88572	72.4	57.5	(57.2, 57.8)	49.4	(49.0, 49.7)
Female	33807	27.6	68.2	(67.7, 68.6)	59.6	(59.0, 60.2)
HIV diagnosis year						
1996	17708	14.5	58.7	(57.9, 59.4)	50.0	(49.2, 50.8)
1997	16564	13.5	59.8	(59.1, 60.5)	51.6	(50.8, 52.5)

1998	15227	12.4	59.7	(58.9, 60.5)	51.6	(50.8, 52.5)
1999	14324	11.7	60.6	(59.8, 61.4)	52.2	(51.3, 53.1)
2000	14923	12.2	60.6	(59.8, 61.3)	51.7	(50.8, 52.5)
2001	14545	11.9	61.7	(60.9, 62.5)	52.3	(51.4, 53.2)
2002	14862	12.1	61.7	(60.9, 62.4)	53.1	(52.2, 54.0)
2003	14226	11.6	61.3	(60.4, 62.0)	55.9	(55.0, 56.7)
Log10(viraload)						
1-2	3570	2.9	60.5	(58.9, 62.1)	53.3	(51.6, 55.1)
3	10750	8.8	69.2	(68.4, 70.1)	61.6	(60.6, 62.6)
4	16297	13.3	72.5	(71.8, 73.2)	62.2	(61.3, 63.0)
5	21232	17.4	51.2	(50.5, 51.9)	38.5	(37.8, 39.2)
6-8	7841	6.4	29.5	(28.5, 30.5)	19.1	(18.2, 20.1)
Unknown	62689	51.23	62.8	(62.4, 63.2)	56.8	(56.4, 57.2)
Total	122379	100	60.4	(60.2, 60.7)	52.2	(51.9, 52.5)

P1: Percentage of 1 year AIDS free survival proportion after HIV diagnosis.

CIP1: 95% confidence interval of P1.

P3: Percentage of 3 year AIDS free survival proportion after HIV diagnosis.

CIP3: 95% confidence interval of P3.

MSM: men who have sex with men.

HC: heterosexual contact with a person at high-risk for or a person with HIV infection or AIDS;

MSM+IDU: men who have sex with men and injection drug user.

AIDS: acquired immunodeficiency syndrome; HIV: human immunodeficiency virus.

First CD4 CNT: First CD4<sup>+</sup> T cell count within 6 months of AIDS diagnosis.

Viraload: HIV RNA copy number.

**Table 5. Model comparison ( Estimated excess hazard ratio)**

	Model1 Grouped Bin.*			Model2 grouped Poi.*			Model3 exacted Poi.*, Individual			Model4 Exacted Poi.*, collapsed		
Df.*	7.60E+04			7.60E+04			9.10E+05			7.60E+04		
Deviance Value/Df*.	56179.50			54500			4.84E+05			5.91E+04		
Parameter	0.74			0.719			5.30E-01			7.80E-01		
	RER	low_rer	hi_rer	rer	low_rer	hi_rer	Rer	low_rer	hi_rer	rer	low_rer	hi_rer
Sex												
Male	1.0											
Female	0.99	0.95	1.03	0.99	0.96	1.03	0.98	0.95	1.02	0.98	0.95	1.02
Age at diagnosis												
13-19	1.0											
20-29	1.95	1.67	2.28	1.96	1.68	2.29	1.86	1.59	2.19	2.00	1.71	2.33
30-39	2.36	2.03	2.75	2.37	2.04	2.77	2.27	1.94	2.65	2.43	2.09	2.83
40-49	3.14	2.70	3.65	3.15	2.70	3.67	3.04	2.60	3.56	3.26	2.80	3.80
50-59	4.27	3.67	4.98	4.24	3.63	4.95	4.22	3.60	4.94	4.53	3.89	5.28
60+	6.27	5.37	7.31	6.00	5.14	7.02	6.42	5.48	7.53	6.93	5.94	8.09
Race/ethnicity												
white	1.0											
Black	1.16	1.13	1.18	1.15	1.13	1.17	1.15	1.13	1.18	1.15	1.13	1.17
Hispanic	1.16	1.13	1.19	1.16	1.13	1.18	1.18	1.15	1.21	1.18	1.15	1.21
Asian	0.94	0.86	1.03	0.94	0.86	1.03	0.95	0.87	1.04	0.95	0.86	1.04
Indian	1.41	1.26	1.58	1.37	1.22	1.55	1.42	1.26	1.59	1.41	1.26	1.59
Transmission category												
MSM	1.0											
IDU Male	1.50	1.46	1.54	1.50	1.46	1.53	1.51	1.48	1.55	1.51	1.48	1.55
IDU Female	1.63	1.56	1.71	1.62	1.55	1.70	1.65	1.57	1.73	1.65	1.58	1.74
MSM+IDU	1.10	1.05	1.14	1.09	1.05	1.13	1.08	1.04	1.13	1.08	1.04	1.13
HC Male	1.01	0.98	1.05	1.01	0.98	1.05	1.01	0.97	1.04	1.01	0.98	1.05
HC Female	1.14	1.09	1.19	1.14	1.08	1.19	1.15	1.09	1.20	1.15	1.10	1.20

Unknown	1.49	1.45	1.53	1.48	1.44	1.52	1.54	1.50	1.58	1.54	1.5	1.58
Diagnosis year												
1996	1.0											
1997	0.87	0.84	0.89	0.87	0.85	0.89	0.86	0.84	0.89	0.86	0.84	0.89
1998	0.86	0.84	0.89	0.86	0.84	0.89	0.86	0.83	0.88	0.86	0.84	0.88
1999	0.84	0.81	0.86	0.84	0.82	0.86	0.84	0.81	0.86	0.84	0.81	0.86
2000	0.86	0.83	0.88	0.86	0.84	0.89	0.86	0.83	0.88	0.86	0.83	0.88
2001	0.83	0.81	0.86	0.83	0.81	0.86	0.83	0.81	0.86	0.83	0.81	0.86
2002	0.78	0.75	0.80	0.78	0.76	0.80	0.77	0.75	0.80	0.77	0.75	0.80
2003	0.64	0.62	0.66	0.64	0.62	0.67	0.69	0.66	0.71	0.69	0.66	0.71
First CD4 CNT												
<50	1.0											
50-99	0.65	0.64	0.67	0.65	0.64	0.67	0.64	0.62	0.65	0.64	0.62	0.65
100-199	0.40	0.39	0.41	0.40	0.39	0.41	0.38	0.37	0.39	0.38	0.38	0.39
>=200	0.42	0.41	0.43	0.42	0.40	0.43	0.40	0.39	0.41	0.40	0.39	0.41
Unknown	0.90	0.88	0.92	0.90	0.88	0.92	0.92	0.90	0.94	0.92	0.90	0.94
Log10(viraload)												
1-2	1.0											
3	2.16	1.86	2.50	2.29	1.96	2.69	2.28	1.95	2.67	2.24	1.92	2.62
4	3.61	3.14	4.15	3.88	3.34	4.51	3.90	3.36	4.53	3.83	3.31	4.44
5	5.57	4.87	6.36	6.05	5.24	7.00	6.13	5.30	7.08	6.01	5.21	6.93
6-8	7.78	6.79	8.90	8.44	7.29	9.76	8.66	7.48	10.02	8.49	7.35	9.80
Unknown	12.34	10.82	14.07	13.45	11.66	15.50	14.16	12.28	16.33	13.87	12.10	16.00

\* Bin.: Binomial; Poi.: Poisson; Df.: Degree of freedom.

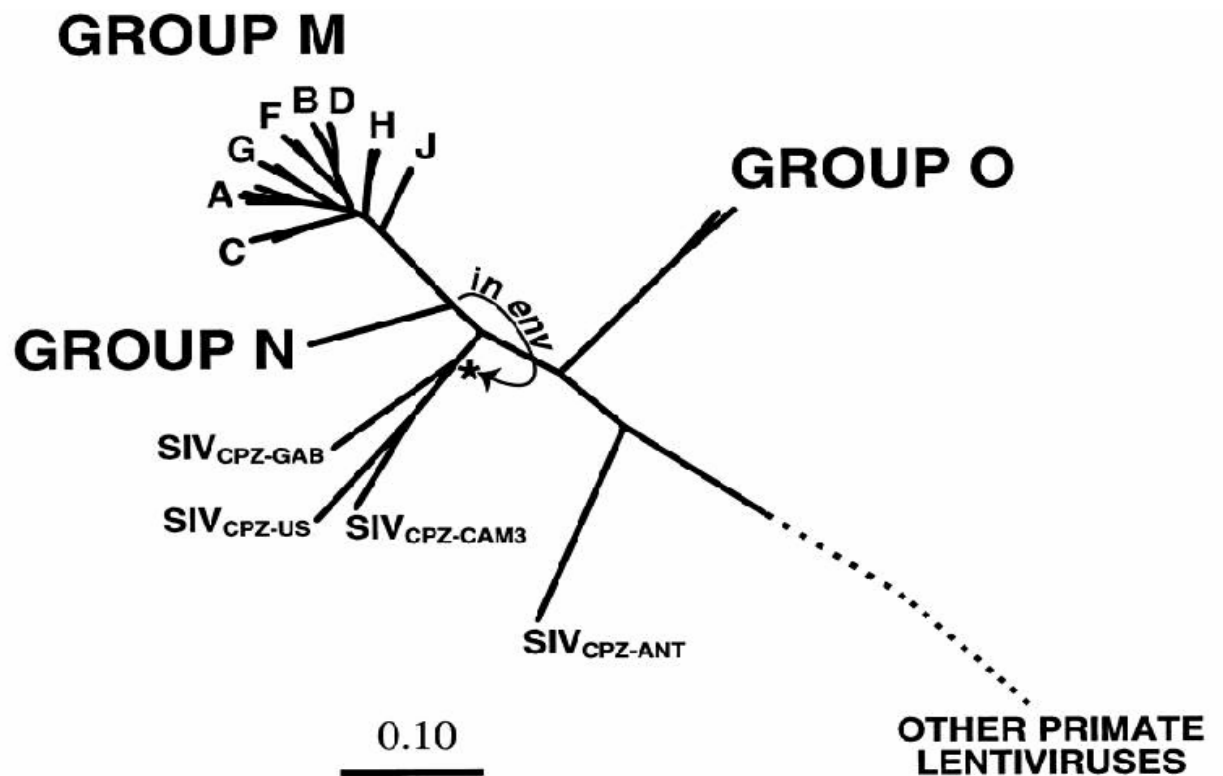
RER: relative excess hazard ratio; low\_rer: low limit of 95% confidence of RER; hi\_rer: upper limit of 95% confidence interval.

HC: heterosexual contact with a person at high-risk for or a person with HIV infection or AIDS;

MSM+IDU: MSM and IDU; AIDS: acquired immunodeficiency syndrome; HIV: human immunodeficiency virus.

First CD4 CNT: First CD4<sup>+</sup> T cell count within 6 months of AIDS diagnosis.

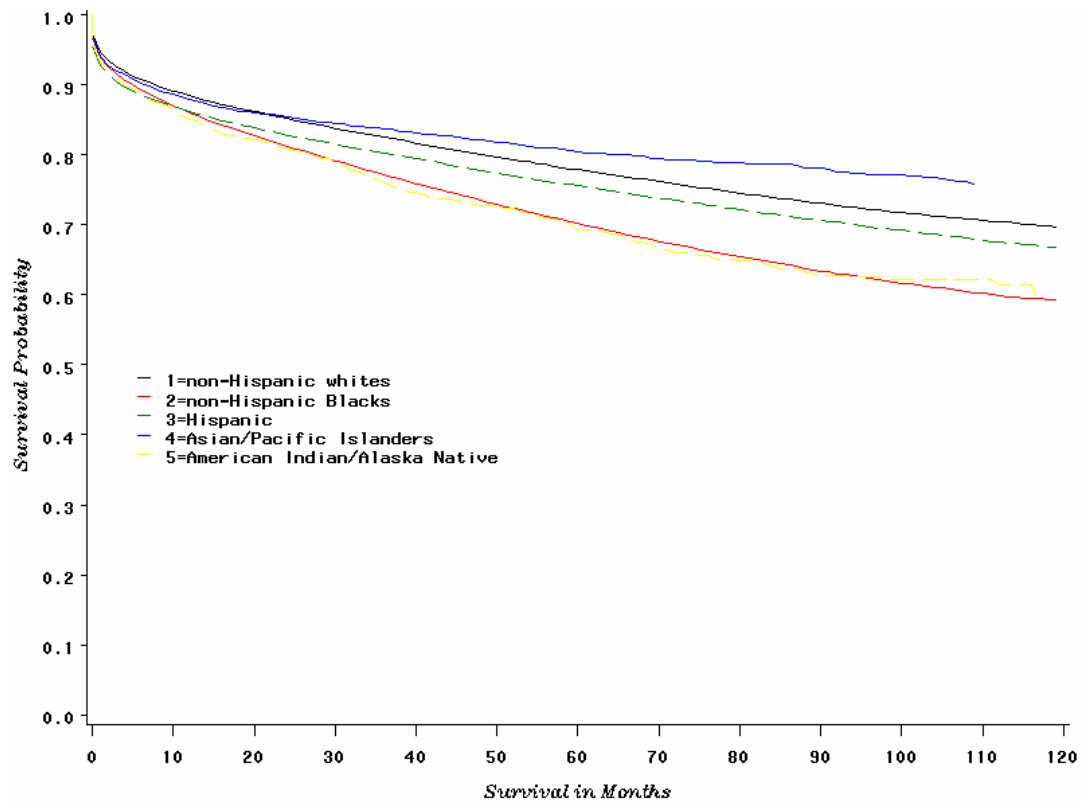
Viraload: HIV RNA copy number.



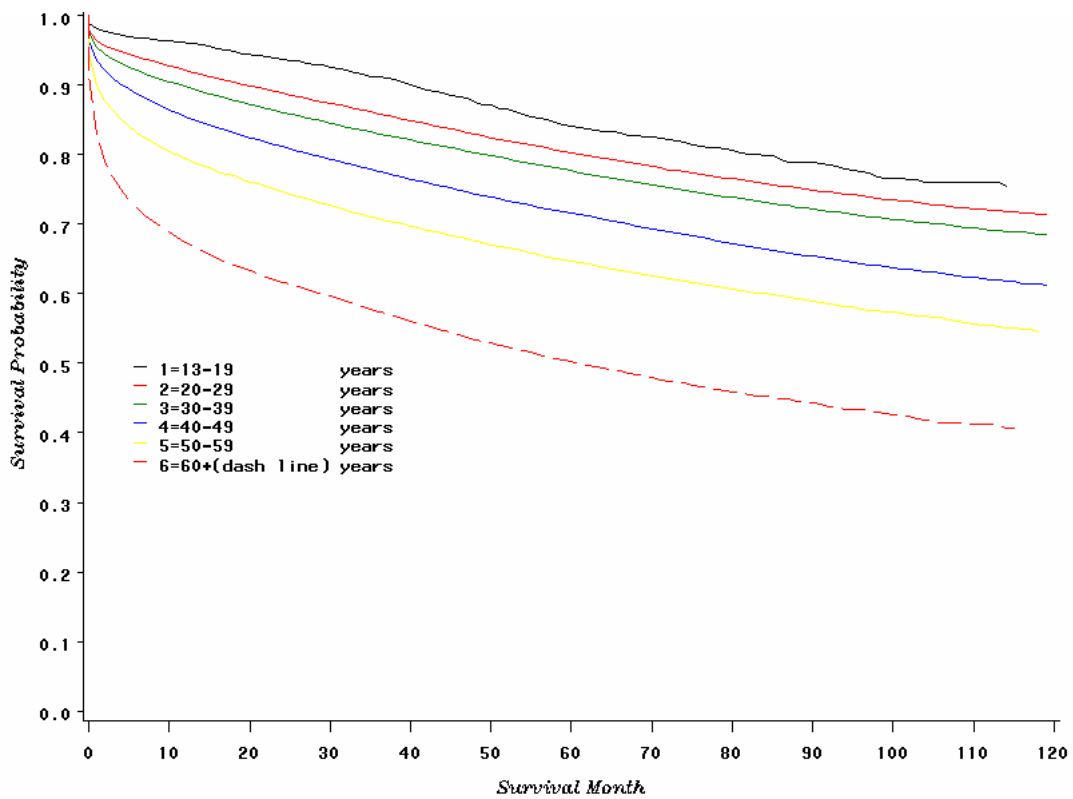
**Figure-1. This maximum likelihood tree of HIV**

based on full-length pol gene sequences analysis, Phylogenetic relationships of HIV-1 groups M, N, and O with four different SIVcpz isolates.

(This picture was from the text book: "Fields Virology, Chapter 60" )

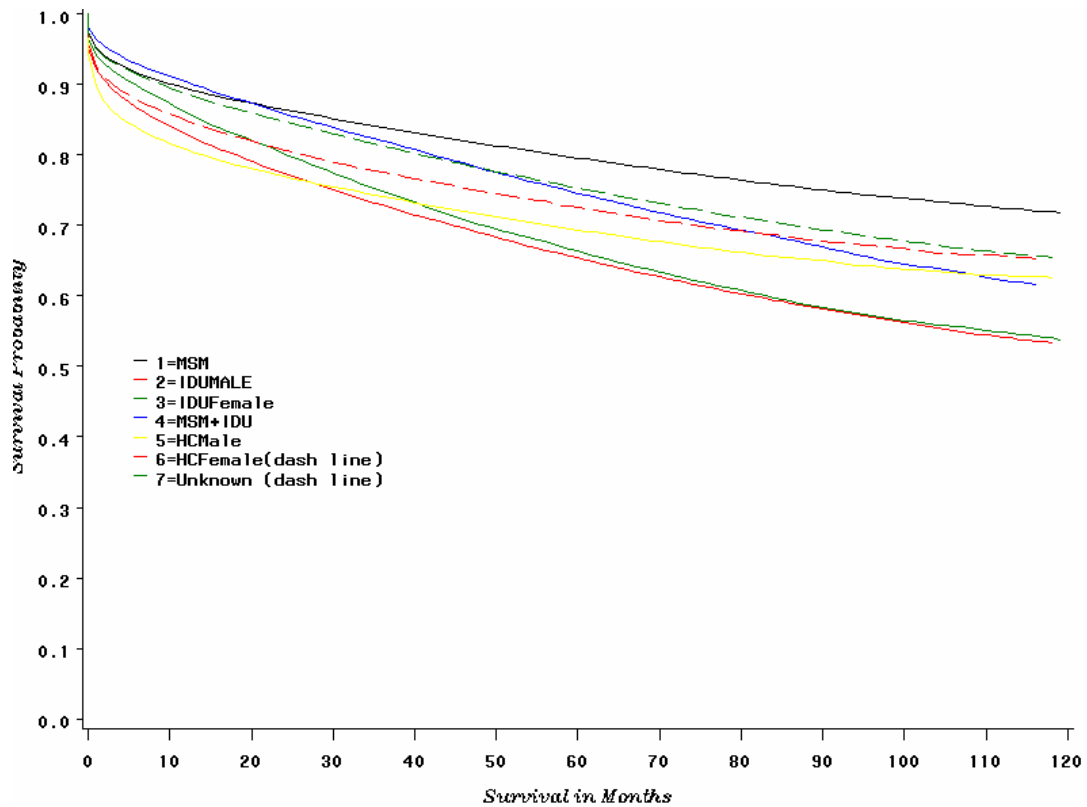


**Figure-2. Survival after AIDS Diagnosis by Race**



**Figure-3. Survival after AIDS Diagnosis by Age Group (years)**



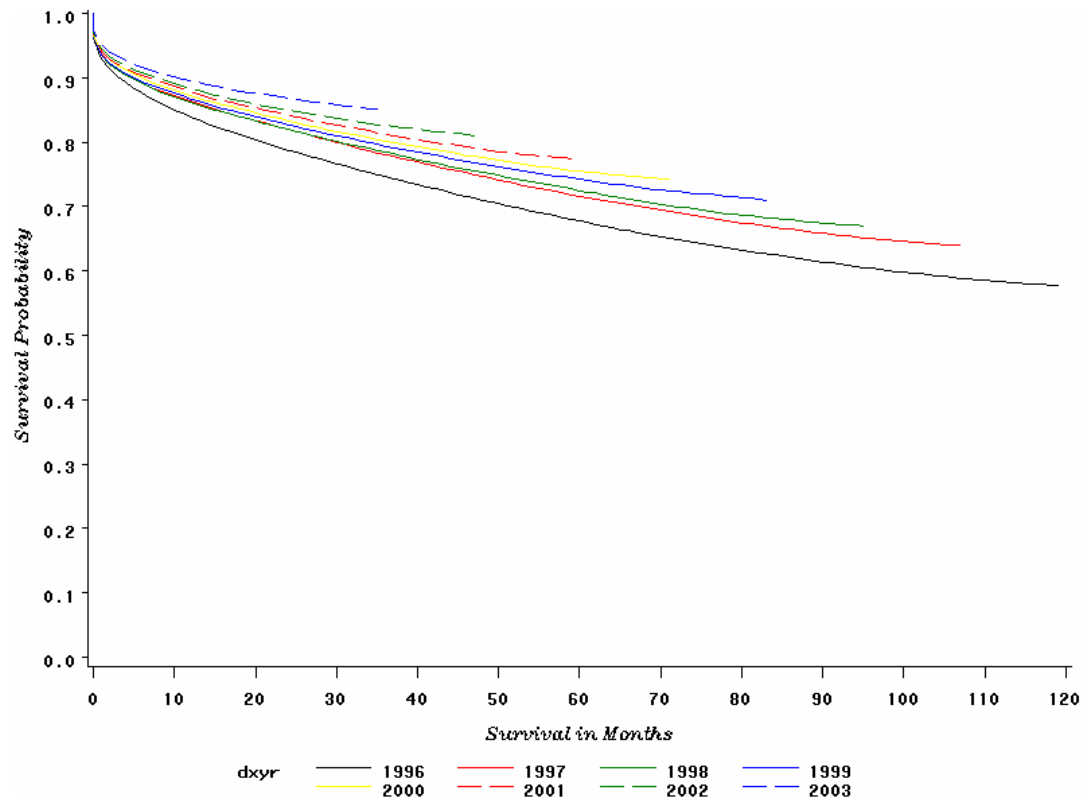


**Figure-4. Survival after AIDS Diagnosis by Transmission Category**

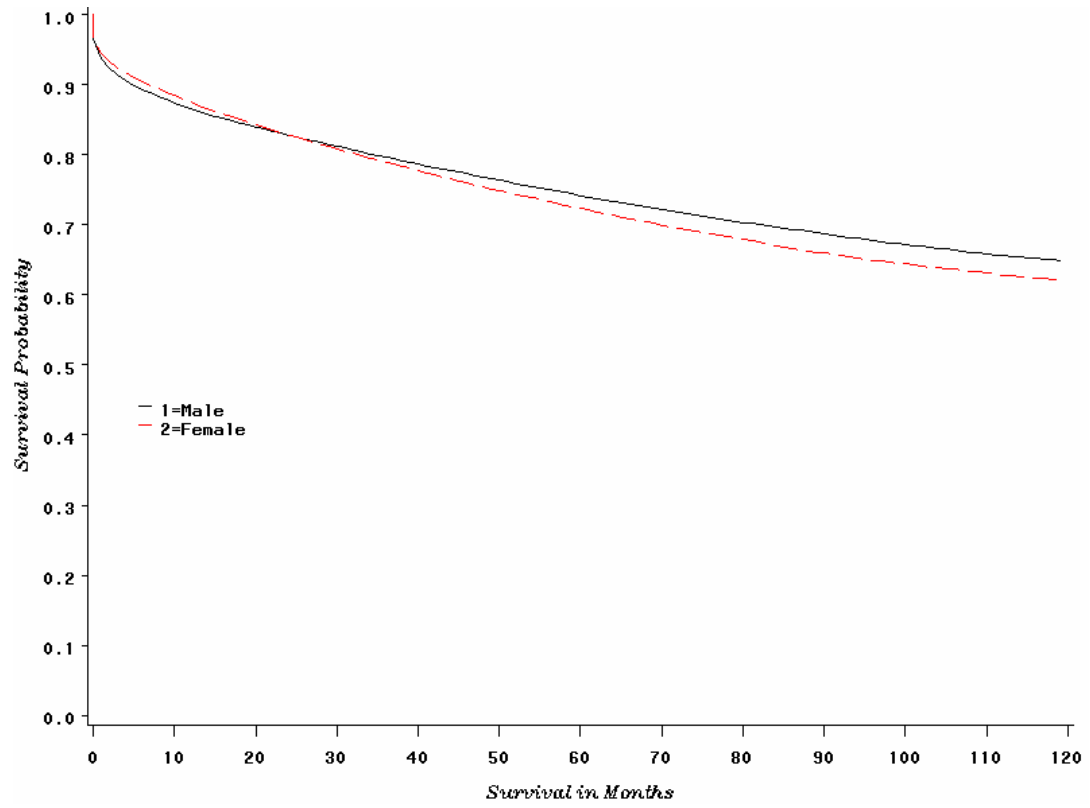
MSM = men who have sex with men; IDU = injection drug user.

MSM+IDU= men who have sex with men and injection drug user.

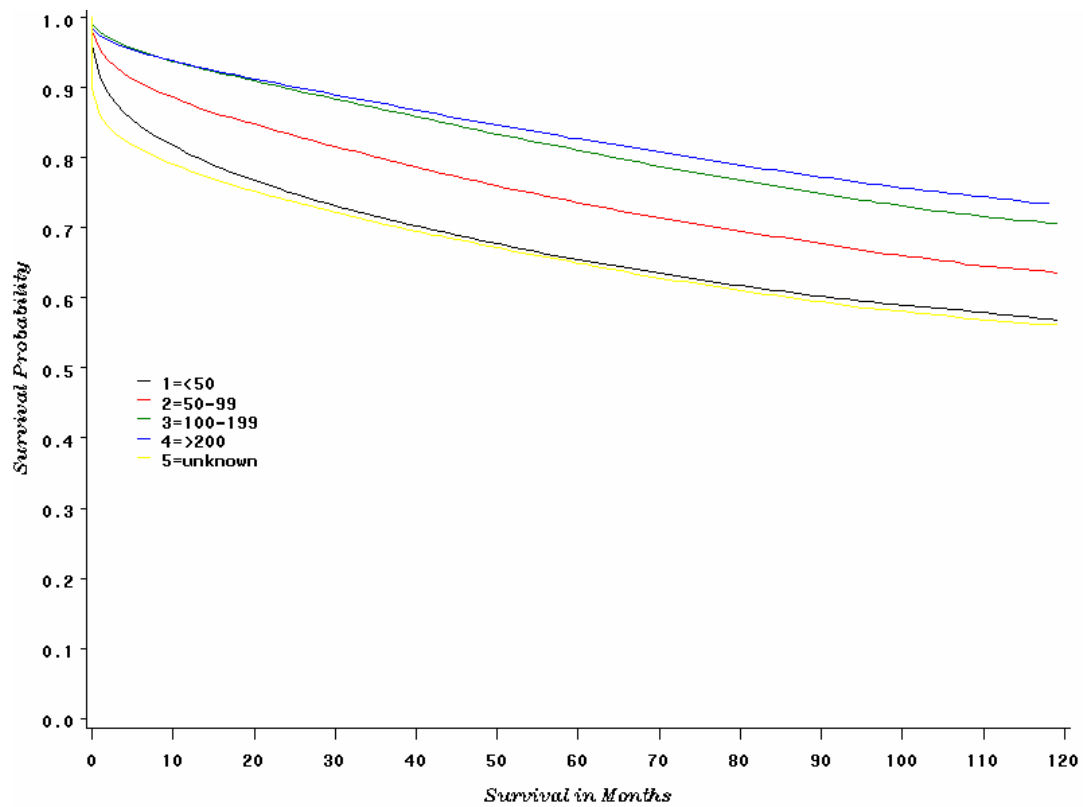
HC = heterosexual contact with a high-risk individual person with HIV infection or AIDS;



**Figure-5. Survival after AIDS Diagnosis by Diagnosis Year**

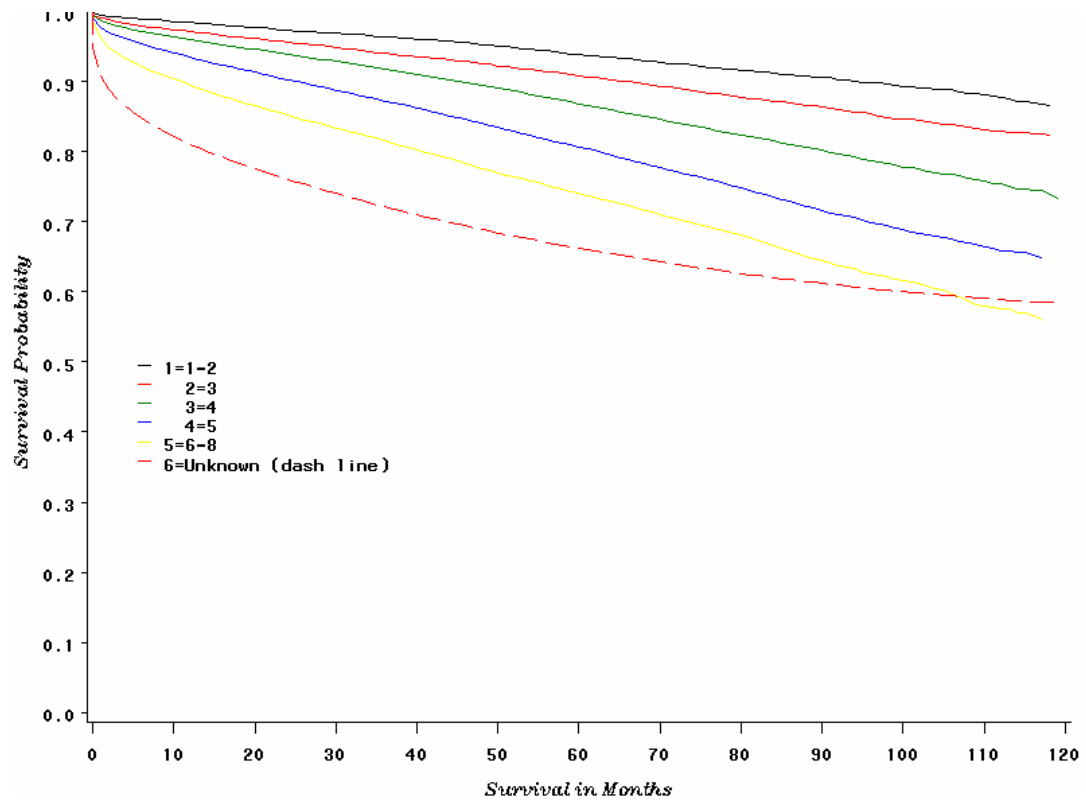


**Figure-6. Survival after AIDS Diagnosis by Sex**



**Figure-7. Survival after AIDS Diagnosis by CD4<sup>+</sup> T Cell Count**

(Note: CD4 count was obtained within 6 months after AIDS diagnosis)



**Figure-8 Survival after AIDS Diagnosis by Viral Load**

Note: The number is Log10 (Viral load); >50% observations had missing value for viral load.

## APPENDICES

### Appendix A

### Abbreviation

AIDS	acquired immune deficiency syndrome
CD4	CD4 molecular positive T lymphocyte
CDC	Centers for Disease Control and Prevention
CI	Confidence interval
CMV	cytomegaloviruses
GLM	generalized linear model
HAART	highly active antiretroviral therapy
HARS	HIV/AIDS reporting system
HC	high-risk heterosexual contact
HCF	female high-risk heterosexual contact
HCM	male high-risk heterosexual contact
HCV	hepatitis C virus
HIV	human immunodeficiency virus
IDU	injection drug user
IDUM	male injection drug user
IDUF	female injection drug user
MSM	men who have sex with men
MSM+IDU	men who have sex with men and inject drugs
RSR	relative survival rate
RR	relative (excess) hazard ratio
Viraload	HIV RNA copy number
RER	Relative excess hazard ratio

## Appendix B SAS and S-plus codes

### 1, SAS code: AIDS to death relative survival analysis ( for table1 )

Note: the main difference of SAS codes for table 2, 3, 5 is the dataset, so I omitted the SAS codes for tables 3, 5

First, to get the AIDS patients dataset, then get the population and mortality dataset and the laboratory dataset. Finally obtain the cohort dataset by merging the following datasets :

- National AIDS patients,
- National population data
- National mortality data,
- Laboratory data of HARS

#### 1.1 SAS codes to obtain the AIDS data

```
options nodate nonumber nocenter ps=65 ls=120;

%let dsmth=JUN06; /* data set month and year */
%let startyr=1996; /* beginning AIDS dx year */
%let enddxyr=2003; /* ending AIDS dx year */
%let endfupyr=2005; /* ending follow-up year */
%let maxfupyr=3; /* maximum follow-up years */run;

title1 "AIDS dx from &startyr to &endyr, reported to CDC as of
&dsmth";

data aids; set hars.as_&dsmth(keep = rep_st stateno race sex
adxyrmo adthyrmo ageaids _amode new_race race_w race_b race_i
race_p race_a hisp);
if ageaids>=13;

dxyr=int(adxyrmo/100);
if &startyr<=dxyr<=&enddxyr;
dxmo=mod(adxyrmo,100);
/* remove cases with missing months of diagnosis */
if dxmo in (0,.) then delete;

dthyr=int(adthyrmo/100);
dthmo=mod(adthyrmo,100);
/* remove cases with missing months of death */
if (dthyr ne .) and ((dthmo in (.,0)) or (adthyrmo<adxyrmo)) then
delete;

if dthyr ne . and dthyr<=&endfupyr
then do; death='1'; survmo=(dthyr*12+dthmo)-(dxyr*12+dxmo); end;
else do; death='0'; survmo=(&endfupyr*12+12)-(dxyr*12+dxmo); end;

*****mapping new race back to old race***;
length old_race $1;
old_race=' ';
if new_race = '1' then old_race = '3';
```

```

else if new_race = '2' then old_race = '5';
else if new_race in ('3','5','7') then old_race = '4';
else if new_race = '4' then old_race = '2';
else if new_race = '6' then old_race = '1';
else if new_race = '8' then do;
    if (race_a = '1' and race_p = '1') and not (race_w = '1' or
race_b = '1' or race_i = '1' or race in('1', '2', '3', '5')) then
old_race = '4';
    else old_race = '9'; end;
else old_race='9';
format old_race $SURRACE. ;
*****;
*****;
if old_race in('1','2','3','4','5');

if _amode = '01' then mode = '1'; /*reclassify risk group*/
else if _amode = '02' then mode = '2';
else if _amode = '03' then mode = '3';
else if _amode = '05' then mode = '5';
else mode='9';

if sex in('1','2');

keep rep_st stateno dxyr dxmo dthyr dthmo death survmo sex
old_race mode ageaids; run;
proc freq data=aids;
tables dxyr dxmo dthyr dthmo death survmo sex old_race mode
ageaids; run;

```

## 1.2 SAS codes to obtain population and mortality data

```

/*get the mortality dataset*/

libname mort '\\cdc\project\DHAP_HICSB_Store1\NCHS MORTALITY
DATA\DthCertData\XUCyyyy'; run;

%macro mort(year=);
proc freq data=mort.Xuc&year;
tables year*sex*old_race*age/out=out&year missing noprint;
run;
%mend;

%mort(year=1996);
%mort(year=1997);
%mort(year=1998);
%mort(year=1999);
%mort(year=2000);
%mort(year=2001);
%mort(year=2002);
%mort(year=2003);
%mort(year=2004);

data mort; set out1996 out1997 out1998 out1999 out2000 out2001
out2002 out2003 out2004;

```



```

death=count;
keep year sex old_race age death; run;

/*get the population data */
%macro pop(range=,year=);
proc freq data=census.cty_&range;
  tables sex*old_race*age/out=out&year missing noprint;
  weight pop&year;
  run;
%mend;

%pop(range=9099,year=96);
%pop(range=9099,year=97);
%pop(range=9099,year=98);
%pop(range=9099,year=99);

%pop(range=0005,year=00);
%pop(range=0005,year=01);
%pop(range=0005,year=02);
%pop(range=0005,year=03);
%pop(range=0005,year=04);
%pop(range=0005,year=05);

data pop; set out96(in=y96) out97(in=y97) out98(in=y98)
out99(in=y99) out00(in=y00)
              out01(in=y01) out02(in=y02) out03(in=y03)
out04(in=y04) out05(in=y05);
  pop=count;
  if y96 then year='1996'; else
  if y97 then year='1997'; else
  if y98 then year='1998'; else
  if y99 then year='1999'; else
  if y00 then year='2000'; else
  if y01 then year='2001'; else
  if y02 then year='2002'; else
  if y03 then year='2003'; else
  if y04 then year='2004'; else
  if y05 then year='2005';
  keep year sex old_race age pop; run;

title 'National population by year, sex, old_race, and age';

proc freq data=pop;
  tables year sex old_race age/missing;
  weight pop; run;

```

## Merge the population and mortality data

```

data mort2005; set mort;
  if year='2004';
  year='2005';
run;
data mort; set n.mort mort2005;

```

```

    if old_race='9' then delete;
    if age>85 then age=85;
proc means noprint data=mort; by year sex old_race age;
    var death;
    output out=mort85 sum=death;

data prob; merge n.pop mort85; by year sex old_race age;
    if age>=13;
    _year=1*year;
    _age=age;
    prob=1-death/pop;
    keep _year sex old_race _age prob; run;

```

### After obtaining the AIDS, population , mortality , laboratory data, merging them

```

/*get the aids dataset*/
data one ;
set a.aids ; by rep_st stateno ;
run;

proc contents data= one; run;

/*get the laboratory dataset */
data two ;
set HIV.lab2; by rep_st stateno; run;

/*get the variables of first cd4 count year and month */
/*remove the cases if their first cd4 count was obtained 6 after
aids diagnosis */
data twol;
set two ; by rep_st stateno;
a= substr(thlyrmo, 1, 2);
b = substr(thlyrmo, 3, 2);
if a in ('.', '..') then delete ;
if b in ('.', '..') then delete ;
    thlyr=int(thlyrmo/100);
    thlmo=mod(thlyrmo,100);
if a= 0 then do _thlyr= 2000 + thlyr; end; *year >= 2000;
else do _thlyr= 1900+ thlyr; end; *year>=1996;
/* remove cases with missing months of first cd4 count test*/
if thlmo in (0,.) then cd4=.;
if 52=<thlyr<=95 then delete; *dxyr between 1952 and 1995;
*if thlyr=6 then delete ; run;

data cases ;
merge one (in = x ) twol; by rep_st stateno;
if x ; run;

/*if the first cd4 count was obtained 6 month after diagnosis,
take the first
cd4 count as missing */
data a.cases1;
set cases; by rep_st stateno;

```

```

if ageaids>=13;
if thlyrmo<adxymrmo then cd4=.;
if _thlyr>= dxyr then do cd4mo =(_thlyr-dxyr)*12 + (thlmo-dxmo);
end;
if cd4mo>6 then do cd4 =.; end ;
if thlyr in (., .. ) then do cd4 =. ; end ;
if thlmo in (., .. ) then do cd4 =. ; end;
run;

*****
/*reclassify the variables */
dataset for frequency table
*****;
data a.cases2;

options nofmterr;
set a.cases1;
race =1*old_race;
agegrp=0;
if 13=<ageaids<=19 then agegrp= 1 ;
else if 20=<ageaids<=29 then agegrp= 2 ;
else if 30=<ageaids<=39 then agegrp= 3 ;
else if 40=<ageaids<=49 then agegrp= 4 ;
else if 50=<ageaids<=59 then agegrp= 5 ;
else if ageaids>=60 then agegrp= 6 ;

_cd4=0;
if 0=<cd4 <50 then _cd4 = 1;
else if 50=<cd4 <=99 then _cd4 = 2;
else if 100=<cd4 <=199 then _cd4 = 3;
else if cd4 >= 200 then _cd4 = 4;
else _cd4 = 5; *frist cd4 count missing;

_virload =0;

if virload in (1 2) then _virload= 1;
else if virload= 3 then _virload = 2;
else if virload= 4 then _virload = 3;
else if virload= 5 then _virload = 4;
else if virload in (6 7 8 ) then _virload = 5; *virload
>6;
else _virload = 6; * virload
missing;

/*reclassify the mode value */

if mode = 5 then mode = 4; *amode = 5 ;
if mode =9 then mode=5; *amode = 9;

if mode =1 then mode =1;
else if mode =2 and sex = 1 then mode = 2;
else if mode = 2 and sex = 2 then mode = 3;
else if mode = 3 then mode = 4;
else if mode = 5 and sex = 1 then mode = 5;

```

```

else if mode = 5 and sex = 2 then mode = 6;
else mode=7;

run;

/* There were 339863 observations read from the data set A.CASES1.
NOTE: The data set A.CASES2 has 339863 observations and 75
variables.
NOTE: DATA statement used (Total process time):*/

data a.aids1; set a.cases2; by rep_st stateno;
race= 1*old_race;
k=min(&maxfupyr,1+int((survmo-0.5)/12));
do fu=1 to k;
  _year=dxyr+fu-1;
  _age=ageaids+fu-1;
  if _age>85 then _age=85;

  if fu<k then do; y=1; d=0; w=0; end; else
  if survmo>36 then do; y=1; d=0; w=1; end;
  else do; y=max(0.5,survmo-(fu-1)*12)/12; d=1*death; w=1-d;
end;
  output;
end;
keep rep_st stateno dxyr fu _year sex _cd4 _virload race _age
ageaids survmo mode y d w; run;

proc sort data=a.aids1; by _year sex race _age; run;

/*There were 913707 observations read from the data set A.AIDS1.
NOTE: The data set A.AIDS1 has 913707 observations and 16
variables.*/

/* import the dataset of population data set containing the
population mortality in a.popmort */

data a.aindivid;
merge a.aids1(in=f) a.popmort;
by _year sex race _age;
if f;
keep rep_st stateno _year sex race _cd4 _virload survmo _age
ageaids dxyr fu mode y d w prob; run;

proc sort data=a.aindivid; by rep_st stateno fu; run;

data a.aindivid1; set a.aindivid;

agegrp=0;
if 13=<_age<=19 then agegrp= 1 ;
else if 20=<_age<=29 then agegrp= 2 ;
else if 30=<_age<=39 then agegrp= 3 ;

```

```

else if 40<=_age<=49 then agegrp= 4 ;
else if 50<=_age<=59 then agegrp= 5 ;
else if _age>=60 then agegrp= 6 ;

length=1;
p_star=prob**length;
d_star=-log(p_star)*(y/length);

label
d_star='Expected number of deaths'
d='Indicator for death during interval'
w='Indicator for censored during interval'
y='Person-time (years) at risk during the interval'
length='Interval length (potential not actual)'
ln_y='ln(person-time at risk)'
p_star='Expected survival probability'
_age='Attained age'
_year='Attained calendar year'
range='Life table interval'
fu='Follow-up interval'
sex='Sex'
;
keep rep_st stateno _year sex race survmo _cd4 _virload agegrp
dxyr fu mode y d w
prob length p_star d_star; run;

proc freq data =a.aindivid1;

tables _cd4 sex race survmo _virload mode agegrp dxyr fu d ;
run;

/*add the variable byvar for computation the survival rate of total
group */
data a.atotal;

options nofmterr;
set a.aindivid1;
byvar=1; run;

proc contents data = a.aindivid1; run;

%macro summary (indata, outdata, var ) ;

proc summary data=&indata nway;
var d w p_star y d_star;
id length;
class &var fu; /* Follow-up must be the last variable in this list
*/
output out=&outdata (drop=_type_ rename=( _freq_=1))
sum(d w y d_star)=d w y d_star mean(p_star)=p_star; run;

%mend summary;

%summary(a.aindivid1, racel, race )
%summary(a.aindivid1, agegrp1, agegrp )
%summary(a.aindivid1, model, mode )

```

```

%summary(a.aindividl, sex1,      sex      )
%summary(a.aindividl, dxyr1,    dxyr    )
%summary(a.aindividl, _cd41,    _cd4     )
%summary(a.aindividl, _virload1, _virload)
%summary(a.atotal,   total1,    byvar   )

*****
*****
/*computation and selection of 1, 3-year cumulative relative
survival probability*/
*****
*****;

%macro surv(indata,outdata, var, num );
data outdata;
retain cp cp_star cr 1;
set &indata;
where fu<=3; * maximum follow up to 10 year, which can be changed
by requirement;
if fu=1 then do;
  cp=1; cp_star=1; cr=1; se_temp=0;
  end;
l_prime=l-w/2;
ns=l_prime-d;
/* Two alternative approaches to estimating interval-specific
survival */
/* Must use the hazard approach for period analysis */
p=exp(-(d/y)*length); /* transforming the hazard */
p=1-d/l_prime; /* actuarial approach */
r=p/p_star;
cp=cp*p;
cp_star=cp_star*p_star;
cr=cp/cp_star;
ln_y_group=log(l_prime-d/2);
ln_y=log(y);
d_star_group=l_prime*(1-p_star);
excess=(d-d_star)/y;
se_p=sqrt(p*(1-p)/l_prime);
se_r=se_p/p_star;
se_temp+d/(l_prime*(l_prime-d)); /* Component of the SE of the
cumulative survival */
se_cp=cp*sqrt(se_temp);
se_cr=se_cp/cp_star;

/* Calculate confidence intervals on the log-hazard scale and back
transform */
/* First for the interval-specific estimates */
if se_p ne 0 then do;
  /* SE on the log-hazard scale using Taylor series approximation
*/
  se_lh_p=sqrt( se_p**2/(p*log(p))**2 );
  /* Confidence limits on the log-hazard scale */
  lo_lh_p=log(-log(p))+1.96*se_lh_p;
  hi_lh_p=log(-log(p))-1.96*se_lh_p;
  /* Confidence limits on the survival scale (observed survival) */
  lo_p=exp(-exp(lo_lh_p));

```

```

hi_p=exp(-exp(hi_lh_p));
/* Confidence limits for the corresponding relative survival rate
*/
lo_r=lo_p/p_star;
hi_r=hi_p/p_star;
/* Drop temporary variables */
drop se_lh_p lo_lh_p hi_lh_p;
/* Formats and labels */
format lo_p hi_p lo_r hi_r 8.5;
label
lo_p='Lower 95% CI for P'
hi_p='Upper 95% CI for P'
lo_r='Lower 95% CI for R'
hi_r='Upper 95% CI for R'
;                               end;

/* Now for the cumulative estimates */
if se_cp ne 0 then do;
/* SE on the log-hazard scale using Taylor series approximation
*/
se_lh_cp=sqrt( se_cp**2/(cp*log(cp))**2 );
/* Confidence limits on the log-hazard scale */
lo_lh_cp=log(-log(cp))+1.96*se_lh_cp;
hi_lh_cp=log(-log(cp))-1.96*se_lh_cp;
/* Confidence limits on the survival scale (observed survival) */
lo_cp=exp(-exp(lo_lh_cp));
hi_cp=exp(-exp(hi_lh_cp));
/* Confidence limits for the corresponding relative survival rate
*/
lo_cr=lo_cp/cp_star;
hi_cr=hi_cp/cp_star;
/* Drop temporary variables */
drop se_lh_cp lo_lh_cp hi_lh_cp;
/* Formats and labels */
format lo_cp hi_cp lo_cr hi_cr 8.5;
label
lo_cp='Lower 95% CI for CP'
hi_cp='Upper 95% CI for CP'
lo_cr='Lower 95% CI for CR'
hi_cr='Upper 95% CI for CR'
;                               end;

drop se_temp;
label
/* range='Interval' */
fu='Interval'
l='Alive at start'
l_prime='Effective number at risk'
ns='Number surviving the interval'
d='Deaths'
w='Withdrawals'
p='Interval-specific observed survival'
cp='Cumulative observed survival'
r='Interval-specific relative survival'
cr='Cumulative relative survival'
p_star='Interval-specific expected survival'
cp_star='Cumulative expected survival'

```

```

ln_y_group='ln(1_prime-d/2)'
ln_y='ln(person-time) (using exact times)'
y='Person-time at risk (using exact times)'
d_star='Expected deaths (using exact times)'
d_star_group='Expected deaths (approximate)'
excess='Empirical excess hazard'
se_p='Standard error of P'
se_r='Standard error of R'
se_cp='Standard error of CP'
se_cr='Standard error of CR'
;
format p cp r cr p_star cp_star 5.3;
run;

proc sort data = outdata; by &var fu ; run;

data outdata;
retain r1 cr1 cr3 cr5 cr10 cr20 cir1 cicr1 cicr3 cicr5 cicr10 ;
length cicr5 $ 12;
set outdata;
by &var fu;

if fu=1 then do; r1=.; cr1=.; cr3=.; cr5=.; cr10=.; end;
if fu=1 then do; r1=r; cir1='(|put(lo_r,4.2)|',
'|left(put(hi_r,4.2))|')';
cr1=cr; cicr1='(|put(lo_cr,4.2)|',
'|left(put(hi_cr,4.2))|')'; end;
if fu=3 then do; r3=r; cir3='(|put(lo_r,4.2)|',
'|left(put(hi_r,4.2))|')';
cr3=cr; cicr3='(|put(lo_cr,4.2)|',
'|left(put(hi_cr,4.2))|')'; end;

*if fu=5 then do; *cr5=cr; *cicr5='(|put(lo_cr,4.2)|',
'|left(put(hi_cr,4.2))|')'; *end;
*if fu=10 then do; *cr10=cr; *cicr10='(|put(lo_cr,4.2)|',
'|left(put(hi_cr,4.2))|')'; *end;
*if fu=20 then do; *cr20=cr; *cicr20='(|put(lo_cr,4.2)|',
'|left(put(hi_cr,4.2))|')'; *end;

keep &var r1 cir1 cr1 cicr1 r3 cir3 cr3 cicr3;

label
l_zero='l_zero'
r1='1-year RSR'
r3='3_year RSR'
cr1='1-year cum RSR'
cr3='3-year cum RSR';
*cr5='5-year cum RSR'
*cr10='10-year cum RSR';
*cr20='20-year cum RSR';
run;

data &outdata;

set outdata ;
do i = 1 to &num;

```



```

m=3*i;
if _n_ = m then output;
end;

keep &var fu r1 cir1 cr1 cicr1 r3 cir3 cr3 cicr3; run;
%mend;

%surv(race1,      race2,      race,      5 );
%surv(agegrp1,   agegrp2 ,   agegrp,   6 )
%surv(mode1,    mode2,      mode,      7 )
%surv(sex1,     sex2,       sex,       2 )
%surv(dxyr1,    dxyr2,     dxyr,     8 )
%surv(_cd41,    _cd42 ,    _cd4,     5 )
%surv( _virload1, _virload2, _virload, 6 );
%surv(total1,   total2,    byvar,    1 );

/*get the frequency for the aids relative survival table */

%Macro freq(indata, var1, var2, var3, var4, var5, var6, var7 );
    %do i=1 %to 7;
        ods output Freq.table&&i..OneWayFreqs=tab&&i;
    %end;
    proc freq data = &indata;
        tables  race agegrp mode sex dxyr _cd4 _virload; run;
%Mend freq;

%freq (a.cases2, race, agegrp, mode, sex, dxyr _cd4 _virload )

%macro clean(indata, outdata, var );

    data &outdata;
    set &indata (keep = &var frequency percent );
    run;

%mend clean ;

%clean(tab1, race3, race )
%clean(tab2, agegrp3, agegrp )
%clean(tab3, mode3, mode )
%clean(tab4, sex3, sex )
%clean(tab5, dxyr3, dxyr )
%clean(tab6, _cd43, _cd4)
%clean(tab7,_virload3, _virload )

/*merge the frequency table and the relative survival table
for each variables */

%macro merge(indata1, indata2, outdata, var );
    data &outdata ;
    merge &indata1 &indata2; by &var;
    run;
%mend;

%merge ( race3,      race2,      race4,      race )
%merge ( agegrp3,   agegrp2,    agegrp4 ,   agegrp )

```

```

%merge ( mode3,          mode2,          mode4,          mode   )
%merge ( sex3,           sex2,           sex4,           sex    )
%merge ( dxyr3,         dxyr2 ,         dxyr4,         dxyr   )
%merge ( _cd43,         _cd42,         _cd44,         _cd4   )
%merge ( _virload3,    _virload2,    _virload4,    _virload )

```

```

/*generate the survival table aids to death dataset */

```

The SAS code for table generation was omitted.

## 2. SAS code for table 2 (omitted is similar to the codes for Table1, the only difference is table 2 use HIV to death data

## 3. SAS codes for relative excess hazard analysis (table3 )

### 1, To obtain the effects of AIDS (table 3 column 2,3 )

```

libname aids 'c:\xzhang18';
libname a 'c:\xzhang18\sasdata';

data individ;

options nofmterr;
set a.individ;
run;
data a.individ;
set individ;
ln_y=log(y);
run;

/*****
Collapse the data to produce the grouped .
*****/

proc summary data=a.individ nway;
var d w p_star y d_star;
id length;
class race agegrp sex _cd4 _virload mode dxyr fu; /*Follow-up must
be
the last variable in this list*/
output out=grouped (drop=_type_ rename=( _freq_=1))
sum(d w y d_star)=d w y d_star mean(p_star)=p_star;
run;

data a.grouped;

set grouped; run;

/*****
Calculate life table quantities.
*****/

```

```

data a.grouped;

set a.grouped;

l_prime=l-w/2;
ns=l_prime-d;
/* Two alternative approaches to estimating interval-specific
survival */
/* Must use the hazard approach for period analysis */
p=exp(-(d/y)*length); /* transforming the hazard */
p=1-d/l_prime; /* actuarial approach */

ln_y_group=log(l_prime-d/2);
ln_y=log(y);
d_star_group=l_prime*(1-p_star);
excess=(d-d_star)/y;

label

fu='Interval'
l='Alive at start'
l_prime='Effective number at risk'
ns='Number surviving the interval'
d='Deaths'
w='Withdrawals'
p='Interval-specific observed survival'

p_star='Interval-specific expected survival'

ln_y_group='ln(l_prime-d/2)'
ln_y='ln(person-time) (using exact times)'
y='Person-time at risk (using exact times)'
d_star='Expected deaths (using exact times)'
d_star_group='Expected deaths (approximate)'
excess='Empirical excess hazard'
;
run;

data grouped2;

set a.grouped(obs= 20000) ;
run;

proc freq data = a.grouped ;
tables fu sex race _cd4 _virload mode agegrp dxyr ; run;

ods output parameterestimates=parmest /* parameter estimates */
             modelinfo=modelinfo      /* Model information */
             modelfit=modelfit         /* Model fit information */
             convergencestatus=converge /* Whether the model
converged */
             type3=type3estimates;     /* Type III estimates */

```

```

proc genmod data=a.grouped ;
title 'Poisson error model fitted to collapsed data (based on exact
survival times) [model 4] ';
fwdlink link = log(_MEAN_-d_star);
invlink ilink= exp(_XBETA_)+d_star;
class sex agegrp race mode dxyr _cd4 _virload fu /param= ref
ref=first order= internal ;
model d = fu sex agegrp race mode dxyr _cd4 _virload
fu*sex fu*agegrp fu*race fu*mode fu*dxyr fu*_cd4 fu*_virload
sex*agegrp sex*race sex*dxyr sex*_cd4 sex*_virload
agegrp*race agegrp*mode agegrp*dxyr agegrp*_cd4 agegrp*_virload
mode*dxyr mode*_cd4 mode*_virload
dxyr*_cd4 dxyr*_virload
_cd4*_virload /error=poisson offset=ln_y type3;
run;

```

```
ods output close;
```

```

data parmest;
set parmest;
if df gt 0 then do;
rer=exp(estimate);
low_rer=exp(estimate-1.96*stderr);
hi_rer=exp(estimate+1.96*stderr);
end;
run;

```

```
proc contents data = parmest; run;
```

```
proc contents data= parmest; run;
```

```

data aidspar;
set parmest;
if parameter not in ( 'fu', 'intercept', 'scale' ) ;
aidsrer =rer;
cia='(|'||put(low_rer,4.2)||', '|left(put(hi_rer,4.2))||')';
keep parameter aidsrer CIA;
format aidsrer 4.2 ;
run;

```

2, To obtain the effects of HIV (table 4 column 4,5 )

```

libname aids 'c:\xzhang18';
libname a 'c:\xzhang18';
libname HIV 'G:\hivdata';

```

```

data Hindivid;
set HIV.Hindivid;
run;

```

```
proc contents data = hindivid; run;
```

```

proc freq data= hindivid;
tables      sex race _cd4 _virload  mode dxyr agegrp;
run;
/*****
Collapse the data to produce the grouped .
*****/

proc summary data=Hindivid  nway;
var d w p_star y d_star;
id length;
class race agegrp sex _cd4 _virload mode dxyr fu; /*Follow-up must
be
the last variable in this list*/
output out=grouped (drop=_type_ rename=( _freq_=l))
sum(d w y d_star)=d w y d_star mean(p_star)=p_star;
run;

data HIV.grouped;

set grouped; run;

/*****
Calculate life table quantities.
*****/
data grouped;

set HIV.grouped;

l_prime=l-w/2;
ns=l_prime-d;
/* Two alternative approaches to estimating interval-specific
survival */
/* Must use the hazard approach for period analysis */
p=exp(-(d/y)*length); /* transforming the hazard */
p=1-d/l_prime; /* actuarial approach */

ln_y_group=log(l_prime-d/2);
ln_y=log(y);
d_star_group=l_prime*(1-p_star);
excess=(d-d_star)/y;

label

fu='Interval'
l='Alive at start'
l_prime='Effective number at risk'
ns='Number surviving the interval'
d='Deaths'
w='Withdrawals'
p='Interval-specific observed survival'

p_star='Interval-specific expected survival'

ln_y_group='ln(l_prime-d/2)'

```

```

ln_y='ln(person-time) (using exact times)'
y='Person-time at risk (using exact times)'
d_star='Expected deaths (using exact times)'
d_star_group='Expected deaths (approximate)'
excess='Empirical excess hazard'
;
run;

proc freq data = grouped ;
tables fu sex race _cd4 _virload mode agegrp dxyr ; run;

ods output parameterestimates=parmest /* parameter estimates */
            modelinfo=modelinfo      /* Model information */
            modelfit=modelfit        /* Model fit information */
            convergencestatus=converge /* Whether the model
converged */
            type3=type3estimates;     /* Type III estimates */

proc genmod data=grouped ;
title 'Poisson error model fitted to collapsed data (based on exact
survival times) [model 4] ';
fwdlink link = log(_MEAN_-d_star);
invlink ilink= exp(_XBETA_)+d_star;
class sex agegrp race mode dxyr _cd4 _virload fu /param= ref
ref=first order= internal ;
model d = fu sex agegrp race mode dxyr _cd4 _virload /
error=poisson offset=ln_y type3;
run;

ods output close;

/*

NOTE: Algorithm converged.
NOTE: The scale parameter was held fixed.
NOTE: The data set WORK.TYPE3ESTIMATES has 8 observations and 4
variables.
NOTE: The data set WORK.CONVERGE has 1 observations and 2 variables.
NOTE: The data set WORK.MODELFIT has 5 observations and 4 variables.
NOTE: The data set WORK.MODELINFO has 7 observations and 6
variables.
NOTE: The data set WORK.PARMEST has 36 observations and 9 variables.
NOTE: PROCEDURE GENMOD used (Total process time):
            real time           3:52:16.44
            cpu time            3:51:55.22

*/

data parmest;
set parmest;
if df gt 0 then do;
rer=exp(estimate);
low_rer=exp(estimate-1.96*stderr);
hi_rer=exp(estimate+1.96*stderr);

```

```

end;
run;

proc contents data = parcest; run;

proc contents data= parcest; run;

data HIV.HIVpar;
set parcest;
if parameter not in ( 'fu', 'intercept', 'scale' ) ;
HIVrer =rer;
ciHIV='('||put(low_rer,4.2)||', '||left(put(hi_rer,4.2))||)';
keep parameter HIVrer CIHIV;
format HIVrer 4.2 ;
run;

data hiv.modelfit;

set modelfit; run;

data hiv.type3estimates;
set type3estimates; run;

```

#### 4. SAS codes for Table 4(omitted ), the SAS code difference between table1 and 4 is Table 4 use HIV to AIDS data

#### 5. SAS codes for model comparison ( omitted )

#### 6 S-plus codes for mean residual life ( from Dr. Qin )

```

# S code for EL based CI for MRL at xx

alpha<-0.05          # confidence level

# Please input xx, z and xdelta

xx<-12               # a time point at which the MRL needs to be
  estimated
#x<-rexp(n=30, rate=1) # lifetime
#y<- rexp(n=30, rate=0.2) # censoring time

z<-h1[,3]           # vector of min(x,y), observed time
n<-length(z)
xdelta<-h1[,7]
ydelta<-1-xdelta
cenp<-mean(ydelta)  # censoring rate

xkm<-survfit(Surv(z, xdelta)~1, type="kaplan-meier")
xtable<-summary(xkm)
xt<-xtable$time     # distint death time from x sample
xtx<-xt[xt>=xx]    # distint death time greater than xx

```

```

xtx1<-c(xt[xt<xx][length(xt[xt<xx])],xtx[-length(xtx)])
Fhat<-xtable$surv          # The K-M estimate for 1-F
Fhatx<-Fhat[xt>=xx]
Fhatx1<-c(Fhat[xt<xx][length(Fhat[xt<xx])],Fhatx[-length(Fhatx)])
dj<-xtable$n.event        # number of deaths at distinct death time
djsx<-dj[xt>=xx]         # number of deaths at death time greater than
                        xx
yj<-xtable$n.risk
yjsx<-yj[xt>=xx]
cmu<-cumsum(Fhatx1*(xtx-xtx1))
meanmu<-cmu[length(cmu)]  # meanmu=sum(Fhatx1*(xtx-xtx1))
sigmax<-sum( ((meanmu-cmu)^2*djx/(yjsx-djx))[yjsx>djx] )/((Fhatx1[1])^2)
                        # variance of \wh M(xx)
Mhatxx<-meanmu/Fhatx1[1]
                        # Estimate for the MRL at xx, \wh
S(xx)=Fhatx1[1]

# Normal Approx. based CI for M(xx):

lowmx<-Mhatxx- qnorm(1-alpha/2)*(sigmax)^(1/2)
                        # lower limit of the normal approx. based CI
for M(x)
uppmx<-Mhatxx+ qnorm(1-alpha/2)*(sigmax)^(1/2)
                        # upper limit of the normal approx. based CI
for M(x)
widmx<-uppmx-lowmx
                        # length of normal approximation based CI for
M(xx)

sink("mrlexamp3eout")
cat("Sample size=", n, "level=",1-alpha, "xx=", xx, "Estimated Mxx=",
    Mhatxx, "censoring rate=", cenp, "\n")
cat(" ", "\n")

cat("lower limit of the normal approx. based CI for M(x) is:", lowmx, "\n")
cat("upper limit of the normal approx. based CI for M(x) is:", uppmx, "\n")
cat("length of the normal approx. CI for M(x) is:", widmx, "\n")

sink()

Mhatxx

Lowmx

uppmx

```