The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 815-822

Estimation of Seroconversion Dates of HIV by Imputation Based on Regression Model[1)]

Seungyeoun Lee[2)]

Abstract

The aim of this study is to estimate the seroconversion date of the human immunodeficiency virus(HIV) infection for the HIV infected patients in Korea. Data are collected from two cohorts. The first cohort is a group of "seroprevalent" patients who were seropositive and AIDS-free at entry. The other is a group of "seroincident" patients who were initially seronegative but later converted to HIV antibody-positive. The seroconversion dates of the seroincident cohort are available while those of the seroprevalent cohort are not. Estimation of seroconversion date is important because it can be used to calculate the incubation period of AIDS which is defined as the elapsed time between the HIV infection and the development of AIDS. In this paper, a Weibull regression model is fitted for the seroincident cohort using information about the elapsed time since seroconversion and the $CD4^+$ cell count. The seroconversion dates for the seroprevalent cohort are imputed on the basis of the marker of maturity of HIV infection : percent of $CD4^+$ cell count.

Keywords : HIV/AIDS, seroconversion, incubation period, Weibull distribution, imputation.

1. 서 론

1980년대에 AIDS 환자가 최초로 발견된 이후로 HIV/AIDS에 대한 수많은 역학적인 연구가 지금까지 지속적으로 이루어져 왔다. 특별히 HIV의 감염경로는 HIV 종류와 역학적으로 널리 퍼져 있는 지역뿐만 아니라 사용하는 항바이러스 투약에 따라서 다양하게 일어날 수 있으므로 HIV의 감염에 관하여 좀 더 깊이 있게 이해하기 위해서는 다양한 인종과 역학적인 집단에 관하여 연구할 필요성이 제기되었다. 지금까지 이러한 연구들은 대부분 미국과 유럽을 중심으로 활발하게 이루어져 왔고, 아시아나 아프리카를 중심으로 하는 연구들은 드물었다. 특별히 한국에서는 HIV/AIDS 환자들에 관한 자료를 많이 얻을 수 없었기 때문에 HIV의 감염경로와 AIDS의 잠복기간에 관한 연구를 진행하기 어려웠다.

---

1) This research was supported by Korea Research Foundation Grant (KRF-2000-041-D00059).
2) Associate Professor, Department of Applied Mathematics, Sejong University, Seoul, 143-747.
E-mail : leesy@sejong.ac.kr

1985년 한국에서 최초로 AIDS 환자가 보고된 이후에 국립보건원에 등록된 HIV/AIDS 환자의 수는 1997년에 765명에 이르렀다. 이 환자들은 등록 이후에도 평균적으로 6개월마다 진단하여 여러 임상적인 자료를 관측하였다. 대부분의 HIV/AIDS 환자의 자료는 HIV에 감염된 시점을 정확하게 알고 있는 경우가 매우 드물다. 이와 같이 환자가 처음으로 관측되는 시점에서 HIV 항체검사결과가 양성반응을 나타내는 환자들을 "seroprevalent cohort"라 하고, 관측시점에서는 HIV 항체검사결과가 음성반응이었다가 관측이 진행되는 과정에서 양성으로 바뀌는 환자들을 "seroincident cohort"라고 한다. 그러나 HIV/AIDS 환자들이 대부분 seroprevalent cohort이므로 HIV 감염시점부터 AIDS가 발병되는 시점까지의 기간으로 정의되는 잠복기간(incubation period)을 정확하게 추정하는데 많은 어려움이 있다. 예를 들어, 이 논문에서 분석하고자 하는 국립보건원에 보고된 자료에서도 seroprevalent cohort의 환자들의 수는 706명인데 반하여 seroincident cohort의 환자들의 수는 59명에 불과하다.

특별히 seroincident cohort에 속한 HIV 감염환자들에 대하여 정확한 감염시기를 알아보기 위하여 HIV 항체검사의 결과에 관한 기록을 조사하였다. 이러한 기록들은 혈액은행이나 검역소 또는 성접촉에 의해 감염되는 질병에 걸린 환자들의 정기적인 HIV 항체검사 결과로부터 얻어진 것이다. 이 환자들에 대하여 seroconversion date는 마지막으로 음성반응이 관측된 시간과 최초로 양성반응이 관측된 시간의 중앙값으로 정의하였다. 그러나 자료의 신뢰도를 높이기 위하여 중앙값이 양성반응이 관측된 시점에서 1년 이내에 관측된 자료를 seroincident cohort로 정의하였다.

HIV 감염과 AIDS의 발병에 관한 자료들은 관측하는 시점에서 이미 HIV에 감염된 환자들이 대부분이므로 HIV에 최초로 감염되는 시점을 알아내는 것이 매우 어렵다. 그러나 AIDS의 잠복기간을 추정하기 위해서는 HIV 감염시점을 알아야 하므로 관측하기 시작한 시점에서 HIV 양성인 환자들이 언제 처음으로 HIV에 감염되었는지를 추정하는 것이 통계학적으로 매우 의미가 크다.

본 논문에서는 $CD4^+$ cell count의 퍼센티지가 HIV의 seroconversion 되는 시점으로부터 경과되는 시간을 예측하는 marker 변수가 되는 것을 이용하여 Weibull 회귀모형을 적합시켰다. 이 모형을 이용하여 seroprevalent cohort의 $CD4^+$ cell ccunt의 퍼센티지에 대한 관측치로부터 seroconversion date를 대치시키는 방법을 제시하고자 한다. 이러한 대치방법에서 고려해야 할 가정은 모형을 적합시키는 모집단에서의 확률분포와 적합시킨 모형을 이용하여 대치값을 구하고자 하는 모집단에서의 확률분포가 동일해야 한다는 것이다. 다시 말하면, 이 경우에 HIV 감염시간의 확률분포가 seroincident cohort와 seroprevalent cohort에서 동일하다는 가정이 필요하다. 또한, 표본의 변동성을 고려하기 위하여 Alcabes, Mu $\tilde{n}$oz and Friedland (1994)에서 시도한 bootstrap 방법을 적용하여 회귀모형을 적합하고 seroconversion date를 대치시키는 다중대치방법(multiple imputation method)을 제안하였다. 국립보건원의 자료를 분석한 결과 HIV의 감염기간의 분포를 파악하기 위하여 seroconversion date의 대치값의 5분위수, 10분위수, 25분위수, 중앙값, 75분위수, 90분위수, 95분위수의 95%신뢰구간을 구하였다.

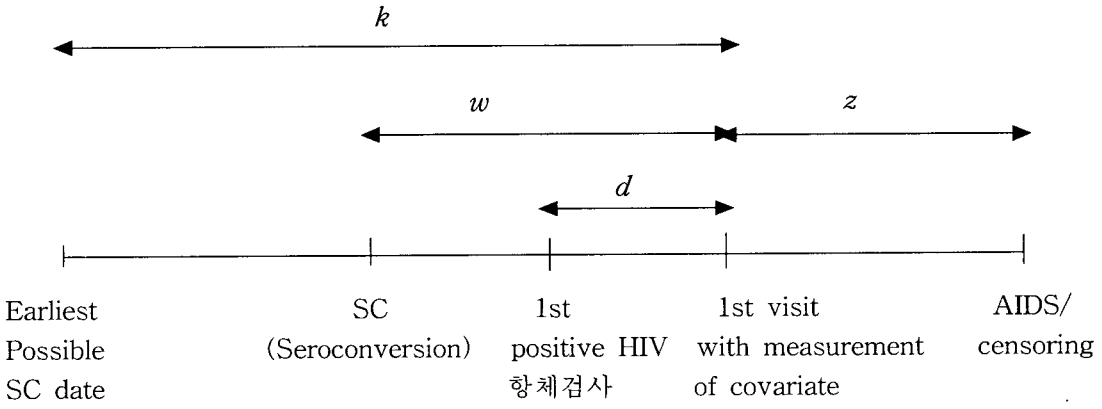
2. Weibull 회귀모형을 이용한 대치방법

어떤 질병의 잠복기간을 정확하게 추정하는 것은 그 질병을 역학적으로 연구하는 측면에서 매우 중요하다. 더욱이 HIV/AIDS와 같이 역학적으로 치명적인 영향을 끼치는 질병에 대하여 잠복

기간을 추정하는 문제는 1980년대에 최초의 발병이후 미국이나 유럽을 중심으로 지속적으로 진행되어 왔다. Mu $\tilde{n}$oz, Carey, Taylor, Chemiel, Kingsley, Raden and Hoover (1992)의 연구에서는 seroprevalent cohort의 HIV감염시점에 관한 추정을 Weibull 모형을 이용하여 시도하였고 Mu $\tilde{n}$oz, Wang, Bass, Taylor, Kingsley, Chemiel and Polk (1989)는 homosexual 남성들의 자료에 대하여 HIV 감염이후부터 AIDS 발병까지의 기간에 대한 추정문제를 다루었다. 또한 Alcabes, Mu $\tilde{n}$oz, Vlahov and Friedland (1994)는 미국의 Baltimore 와 Bronx의 약물중독 AIDS환자들의 자료를 이용하여 HIV 감염기간과 AIDS 잠복기간을 추정하는 문제를 Weibull 회귀모형을 기초로 하여 연구하였다. 본 연구에서는 Alcabes et al. (1994)의 연구에서 가정한 Weibull 회귀모형을 이용하여 HIV 감염기간을 추정하고자 한다.

Weibull 회귀모형을 이용할 수 있는 근거는 HIV에 감염된 시점으로부터 경과되는 시간에 따라 $CD4^+$ cell count의 퍼센티지가 Weibull 모형 하에서 유의하게 감소하는 경향이 나타내기 때문이다. 여기서 $CD4^+$ cell count는 백혈구의 임파구 내에서의 면역성을 나타내는 중요한 marker로서 이 값이 400이하인 경우에는 위험하며 150이하인 경우는 매우 치명적이라고 할 수 있다. 그러나 이 값의 분포가 매우 널리 퍼져 있으므로 자료를 분석할 때에는 백분율로 표시된 $CD4^+$ cell count의 퍼센티지를 이용하였다. 실제로 $CD4^+$ cell count를 주기적으로 관측함으로써 HIV에 감염 예후를 대략 진단할 수 있다.

우선 HIV/AIDS 환자의 자료를 보다 쉽게 이해하기 위하여 Alcabes et al. (1994)에서 나타낸 바와 같이 HIV 감염기간을 (그림 1)과 같이 도식할 수 있다. 여기서 k는 최초로 HIV에 감염될 가능성이 있는 시점부터 처음으로 공변량을 관측하는 시점까지의 시간으로 seroconversion 시점이 취할 수 있는 최대한의 한계이고, d는 HIV 항체검사에서 양성으로 판정된 이후부터 처음으로 공변량을 관측하는 시점까지의 시간으로 seroconversion 시점이 될 수 있는 최소한의 한계이다.



(그림 1)

w 는 seroconversion시점부터 처음으로 공변량을 관측하는 시점까지의 경과시간으로 k 와 d 사이의 값을 취한다. z 는 처음으로 공변량을 관측하는 시점부터 AIDS의 증상이 발병할 때까지

의 시간으로 HIV에 감염이 되었지만 AIDS 증상이 관측이 되지 않는 경우에는 중도절단 되는 시간을 나타낸다.

위의 자료에서 T 를 HIV의 seroconversion 시점으로부터 경과된 시간이라고 정의하고 $CD4^+$ cell count의 퍼센티지와 같은 공변량이 주어졌을 때 T 의 조건부 분포에 대하여 Weibull 회귀모형을 고려해 보자. 그러면 T 의 생존함수는 다음과 같이 주어진다.

$$S(t\ |x) = \exp[-(\gamma t)^{\alpha}]$$

여기서 $\gamma = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$로서 공변량 $x_1, x_2, \cdots, x_p$의 함수이고 $\alpha = \sigma^{-1}$로서 척도모수(scale parameter)의 함수이다. 위의 관계식으로부터 HIV의 seroconversion 시점으로부터 경과된 시간은 다음과 같이 나타낼 수 있다.

$$t = \{-\log[S(w\ |x)]\}^{\sigma} \cdot \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

여기서 $S(w\ |x)$는 seroconversion이 w 보다 전에 일어났을 확률이며 $\beta_0, \beta_1, \cdots, \beta_p$와 σ 는 seroincident cohort에서 추정될 수 있는 모수들이다. 다시 말하면 seroincident cohort에서 Weibull 회귀모형에서 추정된 모수와 생존함수분포를 이용하여 seroprevalent cohort의 seroconversion date를 대치하는 방법을 제안하고자 한다.

Seroprevalent cohort의 환자들의 경우 처음에 양성으로 관측된 시점에서 seroconversion하는 시간이 절단(truncated)되어 관측된 경우이므로 생존함수분포를 추정하고자 할 때 절단된 생존함수분포를 고려하는 것이 필요하다. Seroconversion date의 절단생존함수분포는 다음의 식과 같이 주어진다.

$$S_{tr} = P(w < t < k | d < w < k) = [S(w) - S(k)]/[S(d) - S(k)]$$

여기서 d 는 seroprevalent cohort의 환자들의 HIV 항체검사에서 양성반응결과가 관측된 시점으로부터 처음으로 공변량이 관측된 시점까지의 경과시간으로서 seroconversion이 관측될 수 있는 최소한의 시간을 나타내며, 반면에 k는 seroconversion이 관측될 수 있는 최대시간으로서 실제적으로 AIDS 환자가 최초로 보고된 1985년 9월부터 계산된 시간을 적용하였다. Seroprevalent cohort에 대한 seroconversion date의 대치값을 구하기 위하여 위에서 정의한 절단확률 S_{tr} 가 균등분포(0,1)를 따른다고 가정하면 seroconversion date의 대치값은 다음과 같이 나타낼 수 있다.

$$t = (-\log\{S_{tr}[S(d) - S(k)] + S(k)\})^{\sigma} \cdot \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

실제적으로 Weibull 회귀모형을 이용하여 seroprevalent cohort의 seroconversion 시점을 대치시키는 통계적인 추론과정에서 추정량의 변동을 고려하기 위하여 bootstrap 방법을 적용하였다. 다시 말하면, 회귀모형을 적합하는 과정에서 seroincident cohort에서 표본을 재표본하여 각 표본

마다 회귀계수를 추정하고 이 추정치를 이용하여 seroprevalent cohort에서 재표본한 표본에 대하여 seroconversion date를 대치하는 다중대치법(multiple imputation)을 적용하였다. 이러한 다중대치방법에 bootstrap방법을 응용하는 것은 Alcabes et al.(1994)의 논문에서 적용되었는데 본 논문에서도 이 방법을 적용하였다. 그러나 Alcabes et al.(1994)에서 구한 400개의 대치값보다 더 많은 1000개의 대치값을 구하였다. 이 결과는 seroincident cohort에서는 50개의 표본을 재표본하고 seroprevalent cohort에서는 10개의 표본을 재표본하며 각 표본에 대하여 절단확률을 다르게 2번씩 임의로 추출하여 각 환자에 대하여 1000개의 대치값을 추정하였기 때문이다.

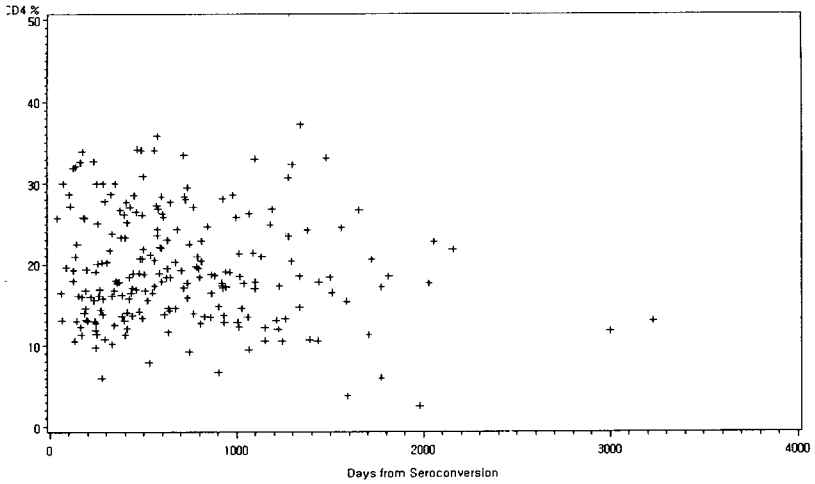
1000개의 표본에서 추정된 대치값을 이용하여 seroprevalent cohort의 seroconversion 시점의 분포를 파악하기 위하여 표본분포의 백분위수 중에서 5분위수, 10분위수, 25분위수, 중앙값, 75분위수, 90분위수와 95분위수를 채택하였다. 각 분위수에서 추정된 1000개의 대치값의 중앙값과 95%신뢰구간을 구하였다. 이 경우 95% 신뢰구간의 하한값은 1000개의 대치값 중 25번째의 순서통계량이고 상한값은 975번째 순서통계량이 된다.

3. 분석 결과

본 논문에서는 국립보건원에서 1985년에 최초로 AIDS 환자가 등록된 이후부터 1997년까지 등록된 765명의 환자에 대한 자료를 바탕으로 HIV 감염기간에 대한 분포를 추정하고자 하였다. 765명의 환자 중에서 59명의 환자들은 HIV 항체검사의 결과가 음성반응에서 양성반응으로 변화된 기록을 보유하고 있으므로 앞에서 언급한 바와 같이 seroincident cohort라고 할 수 있으며, 나머지 706명의 환자들은 관측 초기부터 HIV 항체검사의 결과가 양성반응이었으나 언제 HIV 감염이 일어났는지 정확하게 알 수 없으므로 seroprevalent cohort라고 할 수 있다.

먼저, 59명으로 이루어진 seroincident cohort 자료를 통하여 $CD4^+$ cell count의 퍼센티지가 HIV에 감염된 시점으로부터 경과되는 시간을 예측하는 marker가 되는지를 알아보기 위하여 두 변수간의 산포도를 그려보았다, 실제적으로 이 자료에 있는 환자들은 $CD4^+$ cell count를 비롯한 여러 임상적인 진단자료를 평균 6개월마다 반복적으로 관측하였기 때문에 실제 자료의 수는 59개보다 훨씬 많은 221개이다. 물론 이 자료는 동일한 환자로부터 관측된 반복자료이므로 서로 독립적인 자료라고 볼 수는 없지만 $CD4^+$ cell count의 퍼센티지와 HIV 감염시점으로부터 경과되는 시간과의 산포도를 그릴 때에는 독립성을 고려하지 않고 221개의 자료를 이용하여 (그림 2)를 구하였다. (그림 2)에서 나타난 바와 같이 $CD4^+$ cell count의 퍼센티지는 HIV에 감염된 이후부터 경과되는 시간이 흐름에 따라 약하게나마 감소하는 경향이 있다. 이러한 경향이 통계학적으로 의미가 있는지를 객관적으로 알아보기 위하여 Lognormal, Gamma, Logistic과 Weibull 모형 하에서 회귀모형을 적합하여 본 결과 Weibull 모형에서 유의한 영향을 나타내었다. (표 1)은 모형을 적합한 결과이며 식으로 표시하면 다음과 같다.

$$\log T = 6.1287 - 0.0172\ CD4^+\% + 0.5658\ \varepsilon$$



(그림 2)

(표 1) Seroincident cohort에 대한 Weibull 회귀모형의 적합결과

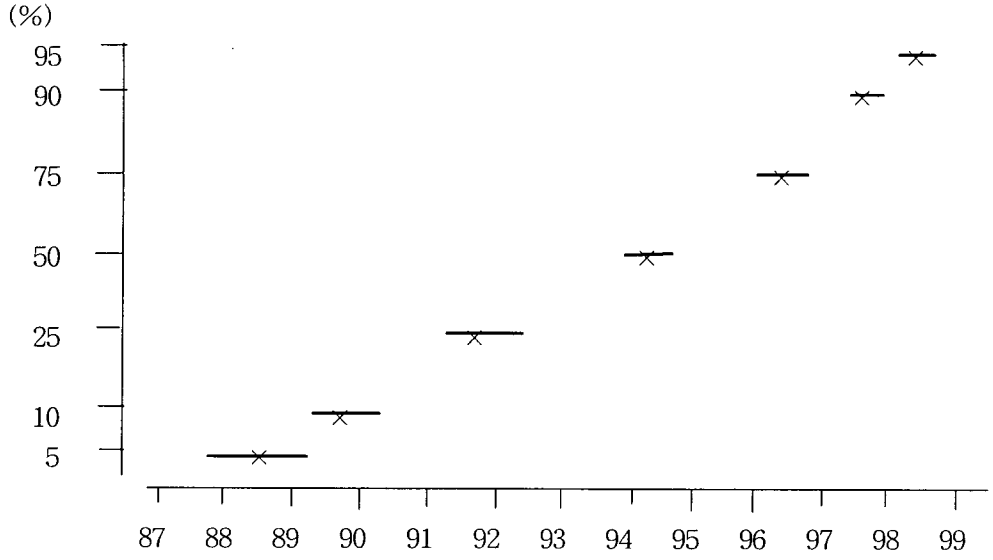
	Estimate (S.E.)	p-value
Intercept	6.1287 (0.177)	0.0001
% of CD4+	-0.0172 (0.0078)	0.0282
Scale	0.5658	
-2 Log likelihood	112.3554	

표 1의 결과를 참조하면 유의수준 0.05에서 $CD4^+$ cell count의 퍼센티지는 HIV 감염이후에 경과하는 시간에 통계학적으로 유의하게 감소하는 경향이 있다고 할 수 있다.

따라서 2절에서 언급한 바와 같이 seroincident cohort의 자료로부터 50개의 표본을 복원추출하여 각 표본에 대하여 위와 같이 Weibull 회귀모형을 적합하여 회귀계수를 구한다. 그리고 균등분포에서 두 번 반복하여 임의로 추출한 절단확률을 이용하여 seroprevalent cohort의 표본에 대하여 seroconversion 시점의 대치값을 구한다. 이 경우에도 추정량의 변동성을 고려하기 위하여 seroprevalent cohort에서 10개의 표본을 복원추출한다. 결과적으로 seroprevalent cohort의 각 환자들에 대하여 1000개의 표본대치값을 구하였다.

이와 같이 다중대치방법으로 구한 seroprevalent cohort의 seroconversion 대치값의 중앙값은 11개월로 추정된다. 다시 말하면 HIV 항체검사의 결과가 양성반응으로 나온 시점에서 1년 이내에 HIV seroconversion 될 가능성이 50%정도라는 것이다.

(그림 3)은 5분위수, 10분위수, 25분위수, 중앙값, 75분위수, 90분위수와 95분위수에서 1000개의 대치값의 중앙값과 95%신뢰구간을 구하여 나타낸 결과이다. (그림 3)을 살펴보면 95% 신뢰구간의 길이가 최근에 관측되어진 자료일수록 짧아지는 것을 알 수 있으며 신뢰구간의 길이는 대부분 1년 이내이다. 또한 이 자료에 관측된 seroprevalent cohort의 50%에 해당하는 환자들은 이 자료가 수집되었던 97년 말을 기준으로 대략 3년 전에 HIV에 감염되었다는 것을 알 수 있다.



(그림 3)

4. 결 론

1980년대에 들어서 AIDS환자가 보고된 이후로 AIDS는 주로 동성과의 성적관계 또는 약물투여와 수혈하는 과정에서 발병하며 10년 가까운 잠복기로 인하여 예방치료가 매우 힘든 불치의 병으로 밝혀져 왔다. 또한 어떤 한 개인이 AIDS 환자로 밝혀지게 되면 그는 사회적으로 고립되는 극한상황을 맞게 되므로 AIDS 환자에 대한 정확한 통계를 알아내는 것 자체가 어렵다. 그러나, AIDS 질병이 국민 건강에 끼치는 지대한 역학적인 영향을 고려해 볼 때 국가적인 차원에서 AIDS 환자에 대한 관리가 시급하다고 생각된다.

미국과 유럽을 중심으로 활발하게 연구된 결과에 의하면 AIDS의 잠복기간이 10년 내외인 것이 밝혀져 있지만 이러한 결과는 인종적, 그룹간에 다소 차이를 보일 수 있기 때문에 우리나라의 자료를 바탕으로 AIDS의 잠복기간을 추정하는 것이 바람직하다고 생각한다. 그러나 AIDS의 잠복기간을 올바로 추정하기 위해서는 HIV 감염시점을 정확하게 관측해야 하는데 대부분의 자료에서는 HIV 감염시점이 절단되어 정확한 잠복기간을 추정할 수가 없다.

본 논문에서는 AIDS의 잠복기를 추정하기 위한 선행연구로서 HIV 감염시점을 추정하는 문제를 다루었다. HIV 감염 이후의 시간이 경과됨에 따라 $CD4^+$ cell count의 퍼센티지가 감소하는 경향이 있으므로 $CD4^+$ cell count의 퍼센티지를 marker로 이용하여 Weibull 회귀모형을 적합하였다. 다시 말하면 seroincident cohort의 자료를 바탕으로 Weibull 회귀모형을 적합시켜 $CD4^+$ cell count의 퍼센티지의 회귀계수를 추정하고 이 추정치를 이용하여 seroprevalent cohort의 자료에서 역으로 HIV 감염시점으로부터의 경과시간을 추정하였다. 이 경우, marker로서 Platelet count와 같은 다른 임상변수도 Weibull 다중회귀모형에서 동시에 고려하였지만 위의 자료에서는 통계학적

으로 전혀 유의하지 않았으므로 $CD4^+$ cell count의 퍼센티지만을 marker로 사용하게 되었다. 표본의 변동성을 고려하여 bootstrap을 이용한 다중대치방법을 응용하여, seroprevalent cohort에 대한 HIV 감염시점을 추정하였다. 이러한 추정결과를 분포함수의 5분위수, 10분위수, 25분위수, 중앙값, 75분위수, 90분위수, 95분위수로 요약하고 각 분위수에 대하여 95%신뢰구간을 구하였다. 추후 연구에서는 추정된 HIV 감염시점을 이용하여 AIDS 잠복기를 추정하고 나아가 AIDS환자의 생존확률을 알아내고자 한다.

References

[1] Alcabes, P., Mu ñoz, A., Vlahov, D., and Frieland, G. (1994). Maturity of HumanImmunodeficiency Virus Infection and Incubation Period of Acquired Immunodeficiency Syndrome in Injecting Drug Users. *Annals of Epidemiology* 4, 17-26.

[2] Kalbfleisch, J.D. and Prentice, R.L. (1980) The Statistical Analysis of Failure Time Data. *John Wiley.*

[3] Mu ñoz, A., Carey, V., Taylor, J., Chmiel, J.S., Kingsley, L.A., Raden, M., and Hoover, D. (1992). Estimation of Time Since Exposure For a Prevalent Cohort. *Statistics in Medicine* 11, 939-952.

[4] Mu ñoz, A., Wang, M.C., Taylor, J., Kingsley, L.A., Chmiel, J.S., and Polk, B.F. (1989). AIDS-free Time After HIV-1 Seroconversion in Homosexual Men. *American Journal of Epidemiology* 130, 530-539.

[5] Taylor, J., Mu ñoz, A., Bass, S.M., Chmiel, J.S., Kingsley, L.A.. and Saah, A.J. (1990). Estimating the Distribution of Times From HIV Seroconversion to AIDS Using Multiple Imputation. *Statistics in Medicine* 9. 505-514.