

Cox 비례위험모형을 이용한 우측 대장암 3기 자료 분석[†]

이태섭¹, 이민정²

^{1,2}강원대학교 정보통계학과

접수 2017년 2월 22일, 수정 2017년 3월 20일, 게재확정 2017년 3월 22일

요약

본 논문에서는 미국 국립암연구소의 SEER 프로그램에서 제공하는 우측 대장암 3기 자료에 Cox 비례위험모형을 적합하여 생존분석을 하였다. 우측 대장암 3기 환자의 사망률에 유의한 영향을 미치는 공변량들을 파악하고, 관심있는 공변량들을 가진 환자의 생존율을 추정하였다. Schoenfeld 잔차를 기반한 검정과 Schoenfeld 잔차 도표, $\log[-\log\{\hat{S}(t)\}]$ 도표를 이용하여 분석에 사용된 공변량들이 비례위험 가정을 만족함을 확인하였다. 적합된 Cox 비례위험모형의 타당성을 검증하기 위해 10-fold 교차 검증을 이용하여 calibration 도표와 시간에 의존하는 ROC 곡선 아래 면적을 계산하였다. 이를 통해 적합된 Cox 비례위험모형의 타당성을 확인하였다.

주요용어: 대장암, 생존분석, 콕스 비례위험모형, calibration 도표, ROC 곡선.

1. 서론

한국인 사망원인 1위는 암으로, 국내 암 발생률은 갑상선암, 위암, 대장암, 폐암, 유방암, 간암, 전립선암의 순으로 높다 (Ministry of Health and Welfare, 2014). Koh와 Kim (2010)은 1999년에서 2007년까지 자료를 바탕으로 대장암 발생률은 연간 6.4%씩 증가하였고 대조적으로 위암과 간암 발생률은 감소하였다고 한다. 국내 대장암 발생률이 증가함에 따라 대장암 연구의 중요성이 부각되고 있으며, 국립암센터와 대한대장항문학회 등에서 다양한 연구가 진행되고 있다.

National Cancer Information Center (2014)에 의하면 대장 중에서 결장과 직장에 생긴 악성종양을 대장암이라 정의한다. 대장의 대부분은 결장이므로 보통 결장암을 대장암으로 쓰기도 한다. 대장암을 나타내는 국제질병분류법 (International Classification of Diseases) 기호는 악성종양의 발생 위치에 따라 정해지며 C18, C19, C20으로 표시하고 C18은 결장, C19는 직장구불결장접합부, C20은 직장에 발생한 경우를 뜻한다 (World Health Organization, 2016). Iacopetta (2002)에 의하면 결장은 근위부 결장(proximal colon)과 원위부 결장 (distal colon)으로 분류되며 이는 흔히 우측 대장과 좌측 대장으로 불린다. 발생학적 측면이나 생화학적인 측면에서 근위부 결장과 원위부 결장은 많은 차이를 가지고 있으므로 이에 따른 악성 종양의 연구도 달라진다. 다른 암과 마찬가지로 대장암의 진행단계는 병기 (stage)로 판정되고 이는 TNM 분류법 (tumor-node-metastasis staging system)과 애슬러-콜러 분류법 (Astler-Coller staging system)을 이용하여 결정한다 (National Cancer Information Center,

[†] 이 논문은 2016년도 정부 (미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2016R1C1B1010294).

¹ (24341) 강원도 춘천시 강원대학길 1, 강원대학교 정보통계학과, 석사과정.

² 교신저자: (24341) 강원도 춘천시 강원대학길 1, 강원대학교 정보통계학과, 조교수.

E-mail: mlee@kangwon.ac.kr

2014). 대장암은 사망률이 높은 폐암에 비해 그 위험도는 비교적 낮으나 국내 대장암 환자수가 크게 증가함에 따라 대장암에 대한 관심도 커져 대장내시경 검사를 건강검진에 포함하는 경우가 많아졌다. 치료과정은 병기에 따라 차이가 있지만 보통 암 종양을 제거하는 수술이 시행되고 있다 (National Cancer Information Center, 2014).

대장암 환자의 사망률과 관련된 요인으로서 환자의 기본정보 뿐만 아니라 대장암 종양의 악성도를 판정하는 기준들을 고려해볼 수 있다. 이러한 기준의 첫 번째로 글리슨 분류법 (gleason grading system)이 있으며 이는 종양 조직 일부를 현미경으로 관찰하여 악성도를 측정한다 (Gleason, 1992). 종양 조직에서 나타나는 선의 형태에 따라 1 등급부터 5 등급까지 등급을 나누는데 주된 형태 (primary pattern)와 보조적 형태 (secondary pattern)에서 두 번 등급을 판정한다. 글리슨 점수 (gleason score)는 주된 형태와 보조적 형태의 등급을 합한 것으로 최소 2 등급에서 최고 10 등급으로 나타낸다. 종양의 악성도는 글리슨 점수가 높을수록 높아지고 점수가 8 이상인 경우는 위험한 단계라고 한다. 대장암 종양의 악성도를 판정하는 두 번째 방법으로는 병기를 세분화하여 확인한다. 대장암 3기의 세분화된 병기 (substage)는 IIIA, IIIB, IIIC로 나뉘지며 주위 림프절 전이정도, 장벽 침윤정도, 다른 장기전이 여부를 종합하여 판정되고 그 위험도는 IIIA, IIIB, IIIC 순으로 높다 (National Cancer Information Center, 2014).

미국 국립암연구소 (National Cancer Institute)의 SEER (Surveillance, Epidemiology, and End Results) 프로그램은 1973년부터 미국 암 환자 자료를 축적하여 미국 인구의 28% 정도의 자료를 보유하고 있으며 연구 목적을 위해 암 자료를 제공하고 있다 (National Cancer Institute, 2016). 본 논문에서는 미국 국립암연구소의 SEER 프로그램에서 제공하는 우측 대장암 3기 (stage III proximal colon cancer) 자료를 사용하여 생존모형을 찾고 적합한 생존모형의 타당성을 검증하고자 한다.

본 논문의 구성은 다음과 같다. 2절에는 Cox 비례위험모형의 기본적인 원리와 적합한 모형의 타당성을 검증하는 방법에 대해 설명한다. 3절에서는 우측 대장암 3기 자료에 대한 설명과 그 분석 결과를 보여준다. 4절에서는 결론을 제시하며 마무리한다.

2. 분석 방법

2.1. Cox 비례위험모형

T 를 어떤 사건이 발생하기까지 걸린 시간이라 하고, Z 를 p 개의 공변량들의 벡터라고 하면, 위험함수 (hazard function)는 다음과 같이 정의된다.

$$\lambda(t; Z) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t, Z)}{\Delta t} \quad (2.1)$$

공변량들이 위험함수에 미치는 영향을 알아보기 위해, Cox는 다음과 같은 비례위험모형을 제안하였다 (Cox, 1972).

$$\lambda(t; Z) = \lambda_0(t) \exp(\beta^T Z) \quad (2.2)$$

여기서 $\lambda_0(t)$ 는 미지의 기저위험함수 (baseline hazard function)로서 공변량들의 값들이 모두 영일 때의 위험함수이다. β 는 p 개의 공변량들에 대응하는 회귀계수들의 벡터이다. Cox 비례위험모형에서는 공변량들이 위험함수에 일정한 비율로 변한다. 예를 들어, 두 공변량 $Z = (Z_1, Z_2, \dots, Z_p)^T$ 와 $Z^* = (Z_1^*, Z_2, \dots, Z_p)^T$ 의 위험함수 비는 다음과 같다.

$$\frac{\lambda(t; Z)}{\lambda(t; Z^*)} = \frac{\lambda_0(t) \exp(\beta^T Z)}{\lambda_0(t) \exp(\beta^T Z^*)} = \exp\{\beta_1(Z_1 - Z_1^*)\} \quad (2.3)$$

이는 위험함수 비가 시간에 관계없이 일정함을 보여준다. Cox 비례위험모형의 회귀계수 β 는 부분우도 (partial likelihood)를 최대화 하는 값으로 추정된다. 누적기저위험함수 (cumulative baseline hazard function) $\Lambda_0(t) = \int_0^t h_0(u)du$ 는 Breslow (1974) 추정량을 이용하여 추정될 수 있다.

생존함수 $S(t; Z) = \Pr(T > t|Z)$ 는 식 (2.1)에 주어진 위험함수 정의를 이용하여 다음과 같이 표현될 수 있다.

$$S(t; Z) = e^{-\int_0^t \lambda(u; Z)du} \quad (2.4)$$

Cox 비례위험모형 아래에서는 다음과 같이 표현될 수 있다.

$$S(t; Z) = e^{-\int_0^t \lambda_0(u) \exp(\beta^T Z)du} = e^{-\Lambda_0(t) \exp(\beta^T Z)} \quad (2.5)$$

Breslow (1974) 추정량 $\hat{\Lambda}_0(t)$ 과 최대부분우도추정량 $\hat{\beta}$ 를 이용하여 식 (2.5)에 주어진 생존함수는 다음과 같이 추정된다.

$$\hat{S}(t; Z) = \exp \left\{ -\hat{\Lambda}_0(t) \exp(\hat{\beta}^T Z) \right\} \quad (2.6)$$

2.2. 모형의 타당성 검증

적합된 모형의 타당성은 검정 (calibration)과 판별 (discrimination) 관점에서 평가될 수 있다 (Vergouwe 등, 2002). 적합된 비례위험모형의 검정적 정확성 (calibration accuracy)을 평가하기 위해 calibration 도표를 이용하고, 판별적 정확성 (discriminatory accuracy)을 평가하기 위해 시간에 의존하는 ROC 곡선 아래 면적 (time-dependent area under the ROC curve; AUC(t))을 이용한다. Calibration 도표와 AUC(t)는 10-fold 교차 검증 (cross validation)을 이용하여 계산한다.

2.2.1. 10-fold 교차 검증

경쟁위험자료 (competing risks data) 분석에서 적합된 경쟁위험모형의 타당성을 검증하기 위해 10-fold 교차 검증이 사용되었다 (Lee 등, 2012). 이와 같은 방법은 적합된 Cox 비례위험모형의 타당성을 검증할 때도 사용될 수 있다. 자료를 $n/10$ 크기로 10 등분한 후, 9 등분 자료를 훈련용 자료 (training set)로 사용하고 나머지 1 등분 자료를 시험용 자료 (test set)로 사용한다. 훈련용 자료에 Cox 비례위험모형을 적합하여 추정된 회귀계수를 시험용 자료에 적용하여 시험용 자료 내 모든 개체들에 대해 식 (2.6)과 같이 생존함수를 추정한다. 이와 같은 과정을 다른 등분들이 모두 시험용 자료로 이용될 수 있도록 10번 반복하고, 10개의 시험용 자료로부터 구한 생존함수 추정치들을 저장한다. 10-fold 교차 검증 후에는 자료 내의 모든 개체들이 추정된 생존함수를 가지게 되며, 이 생존함수 추정치들을 이용하여 calibration 도표와 AUC(t)를 계산한다.

2.2.2. Calibration 도표

적합된 Cox 비례위험모형을 통해 예측한 생존율이 관측된 생존율과 얼마나 비슷한 지를 calibration 도표를 이용하여 검증한다 (Miller 등, 1993). 10-fold 교차 검증을 통해 구한 생존함수 추정치들의 사분위수 (quartile)를 구하고, 이를 이용하여 전체 자료를 4 등분한다. 각 등분에서 t 시점에서의 생존함수 $S(t)$ 를 Kaplan-Meier 추정량 (Kaplan과 Meier, 1958)을 이용하여 추정한다. 각 등분에서 10-fold 교차 검증을 이용하여 구한 $\hat{S}(t; Z)$ 들의 평균 $\frac{1}{m} \sum_{j=1}^m \hat{S}(t; Z_j)$ 을 계산한다. 여기서 m 은 4등분된 자료 내 개체들의 수이다. 각 등분의 $\hat{S}(t)$ 와 $\frac{1}{m} \sum_{j=1}^m \hat{S}(t; Z_j)$ 을 쌍으로 나타낸 도표를 calibration 도표라고 부르며, 적합된 모형이 완벽하다면 이 쌍들이 45도 선 위에 위치하게 된다. 이를 통해 적합된 Cox 비례위험모형의 검정적 정확성을 평가할 수 있다.

2.2.3. 시간에 의존하는 ROC 곡선 아래 면적 (AUC(t))

시간에 의존하는 AUC(t)는 t 시점까지 사건을 경험한 개체는 t 시점까지 사건을 경험하지 않은 개체보다 사건을 경험할 가능성이 높을 것이라는 확률이다. 이 확률을 추정하기 위해, 자료 내 모든 개체들을 t 시점까지 사건 발생 유무에 따라 사례군 (case group)과 대조군 (control group)으로 나눈다. 사례군에서 10-fold 교차 검증을 이용하여 구한 t 시점에서의 $\hat{S}(t; Z)$ 를 대조군에서의 t 시점에서의 생존함수 추정치와 비교한다. Heagerty와 Zheng (2005)의 방법을 이용하여 AUC(t)를 계산하였다. AUC(t)가 1에 가깝다면 적합한 모형이 전체 자료를 사례군과 대조군으로 완벽하게 분류할 수 있음을 의미하고, 0.5에 가깝거나 그 이하일 경우는 적합한 모형의 판별적 정확성이 매우 좋지 않음을 의미한다. 이를 통해 적합한 Cox 비례위험모형의 판별적 정확성을 평가할 수 있다.

3. 자료 분석 결과

3.1. 자료 설명

분석에 사용된 자료는 미국 국립암연구소의 SEER 프로그램에서 제공하는 우측 대장암 3기 자료이다. 1996년부터 2005년 사이에 진단받은 40세에서 94세 환자들로 암 중앙 제거 수술을 받았으며, 관측 시간의 중앙값은 34개월이고 최대 120개월까지 추적 (follow-up)하였다. Table 3.1은 분석에 사용된 공변량들의 특성을 보여준다. 18,436명의 우측 대장암 3기 환자 중 6,822명이 암으로 사망하였으며, 나머지 11,614 환자들은 중도절단 (censored) 되었다.

Table 3.1 Data characteristics

Categorical Variables		n	%
Total cases		18,436	
Gender			
	Male	8,173	44.33
	Female	10,263	55.67
Race			
	White	15,166	82.26
	Black	1,803	9.78
	Other	1,467	7.96
Marital status			
	Single	10,333	56.05
	Married	8,103	43.95
Gleason score			
	2-7	12,082	65.53
	8-10	6,354	34.47
Substage			
	Stage IIIA	1,426	7.73
	Stage IIIB	10,917	59.22
	Stage IIIC	6,093	33.05
Vital status			
	Censored	11,614	63.00
	Death from colon cancer	6822	37.00

3.2. Cox 비례위험모형 적합 결과

Table 3.2는 우측 대장암 3기 자료에 Cox 비례위험모형을 적합한 결과를 보여준다. 유의수준 5%에서 성별, 인종, 기혼여부, 글리슨 점수, 세분화된 병기, 나이, 진단년도는 대장암 사망률에 유의한 영향을 미침을 알 수 있다. 공변량들의 특정한 값을 가지는 환자의 생존율을 추정하기 위하여 글리슨 점수가 8-10이고, 세분화된 병기가 IIIB이며, 2005년에 진단받은 기혼의 66세 백인 남자를 선택하였다. Figure

3.1은 이 환자의 추정된 생존율과 95% 신뢰구간을 보여준다. 이 환자가 1년 생존할 확률 (95% 신뢰구간)은 89.97% (89.19%, 90.70%), 5년 생존할 확률은 64.78% (62.61%, 66.86%), 10년 생존할 확률은 57.57% (54.94%, 60.09%)이다.

Table 3.2 Regression parameter estimates, standard errors, hazard ratios, and *p*-values for death from proximal colon cancer in stage III with surgery performed under the Cox proportional hazards model

Covariates		$\hat{\beta}$	se($\hat{\beta}$)	exp($\hat{\beta}$)	<i>p</i> -value
Gender	Male	0	-	1	-
	Female	-0.168	0.026	0.845	<0.001
Race	White	0	-	1	-
	Black	0.231	0.041	1.260	<0.001
	Other	-0.092	0.047	0.912	0.052
Marital status	Married	0	-	1	-
	Single	0.191	0.026	1.211	<0.001
Gleason score	2-7	0	-	1	-
	8-10	0.299	0.025	1.349	<0.001
Substage	Stage IIIA	0	-	1	-
	Stage IIIB	0.777	0.069	2.176	<0.001
	Stage IIIC	1.438	0.070	4.211	<0.001
Age		0.024	0.001	1.024	<0.001
Year of diagnosis		-0.021	0.005	0.979	<0.001

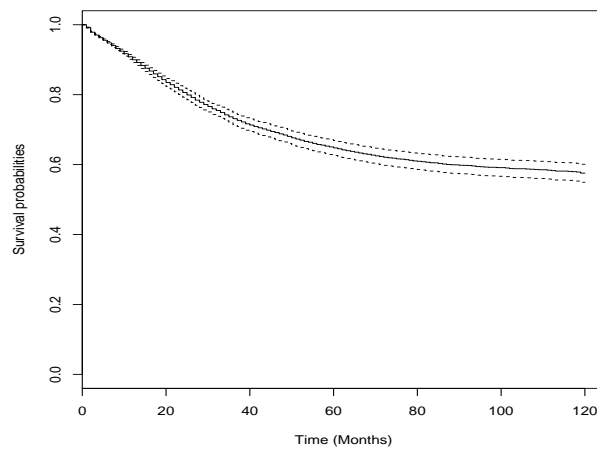


Figure 3.1 Estimated survival probabilities with 95% pointwise confidence intervals for a married white man aged 66 years diagnosed in 2005 with gleason score 8-10 in stage IIIB with surgery performed

3.3. 비례위험 가정 검토

분석에 사용된 공변량들의 비례위험 가정을 검토하기 위해 Schoenfeld 잔차를 기반한 검정 (Grambsch와 Therneau, 1994)과 Schoenfeld 잔차 (Schoenfeld, 1982) 도표, $\log[-\log\{\hat{S}(t)\}]$ 도표를 이용하였다. Table 3.3는 Schoenfeld 잔차에 일반화 선형 회귀모형을 적합하였을 때 그 회귀모형의 기울기가 0이라는 귀무가설에 대한 양측 검정의 *p*-값을 보여준다. 귀무가설이 기각 된다면 해당되는 공변량이 비

레위험 가정을 만족하지 않음을 의미한다. 유의수준 5%에서 성별과 진단년도는 비레위험 가정을 만족함을 알 수 있다. 그러나 나이, 기혼여부, 글리슨 점수, 인종, 세분화된 병기는 기울기가 0이라는 귀무가설을 기각한다. Figure 3.2와 3.3에서 이 공변량들에 대해 Schoenfeld 잔차 도표를 그렸고, 95% 신뢰 구간을 가진 lowess smoothing 곡선도 함께 그렸다. 잔차 도표가 상수가 아닌 다른 패턴을 보인다면 해당되는 공변량은 비레위험 가정을 만족하지 않음을 뜻한다. Figure 3.2와 3.3은 나이, 인종, 세분화된 병기가 비레위험 가정을 만족함을 보여준다. 기혼여부의 잔차 도표에서 아주 미세한 감소추세가 있으며, 글리슨 점수의 잔차 도표도 감소 추세를 보이고 있으므로, $\log[-\log\{\hat{S}(t)\}]$ 도표를 그려 비레위험 가정을 추가적으로 검토하였다.

공변량의 범주에 대해 $\log[-\log\{\hat{S}(t)\}]$ 곡선들이 평행하다면, 그 공변량이 비레위험 가정을 만족함을 뜻한다. 여기서 $\hat{S}(t)$ 는 Kaplan-Meier 추정치이다. Figure 3.4에서 기혼여부와 글리슨 점수의 $\log[-\log\{\hat{S}(t)\}]$ 곡선들이 평행하므로 기혼여부와 글리슨 점수는 비레위험 가정을 만족한다고 할 수 있다. 이를 통해 Cox 비레위험모형에 사용된 모든 공변량들이 비레위험 가정을 만족함을 확인하였다.

Table 3.3 Tests of non-zero slopes in a generalized linear regression of the scaled Schoenfeld residuals

Covariates		Chi-square	<i>p</i> -value
Gender	Male	-	-
	Female	2.062	0.151
Race	White	-	-
	Black	7.513	0.006
	Other	6.506	0.011
Marital status	Married	-	-
	Single	23.615	<0.001
Gleason score	2-7	-	-
	8-10	193.382	<0.001
Substage	Stage IIIA	-	-
	Stage IIIB	8.190	0.004
	Stage IIIC	6.139	0.013
Age		88.491	<0.001
Year of diagnosis		1.158	0.282

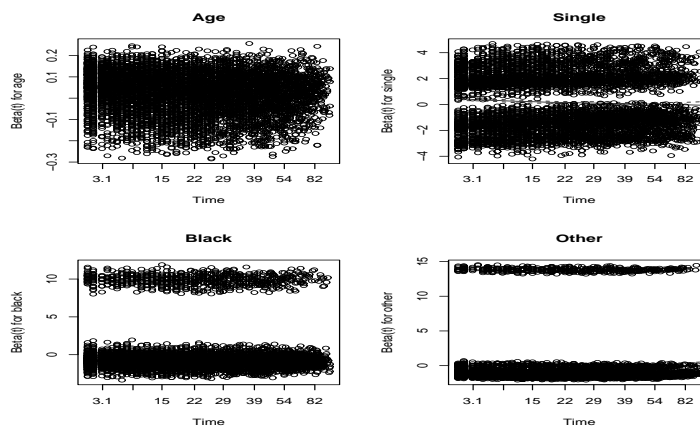


Figure 3.2 Schoenfeld residuals plot for age, marital status, and race

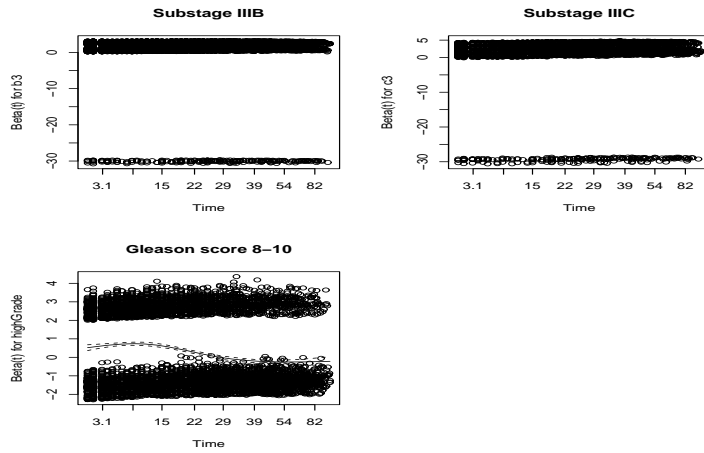


Figure 3.3 Schoenfeld residuals plot for substage and gleason score

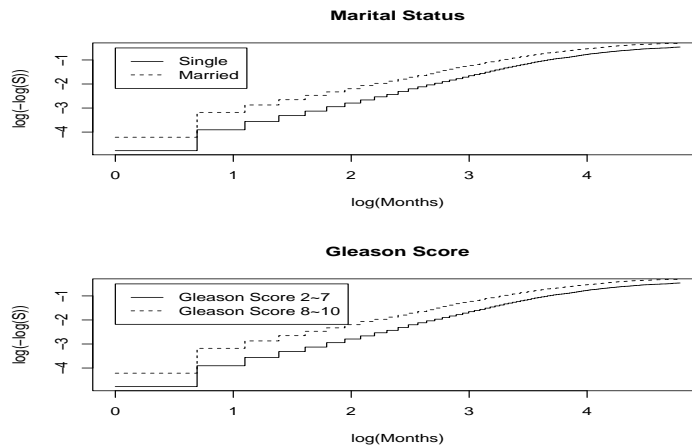


Figure 3.4 $\log[-\log\{\hat{S}(t)\}]$ versus $\log(\text{time})$ for marital status and gleason score

3.4. 모형 타당성 검증 결과

3.4.1. Calibration 도표

Figure 3.5는 우측 대장암 3기 자료의 calibration 도표를 나타낸다. 숫자 1, 3, 5, 7, 10은 생존함수가 예측된 시간 (년)을 나타내고, 알파벳 a, b, c, d 는 예측된 생존함수의 사분위수를 나타낸다. 예를 들면, 점 '5b'와 '10d'는 5년 생존함수 추정치들의 2번째 사분위수와 10년 생존함수 추정치들의 4번째 사분위수를 나타낸다. 그림에서 점선은 45도 선을 나타낸다. 대부분의 점들이 45도 선 위에 있으므로 적합한 비례위험모형에 기반한 생존함수 추정치들의 평균이 Kaplan-Meier 추정치들과 아주 비슷함을 알 수 있다. 이로부터 적합한 Cox 비례위험모형이 대장암 생존율을 예측하기 위한 좋은 모형임을 알 수 있다.

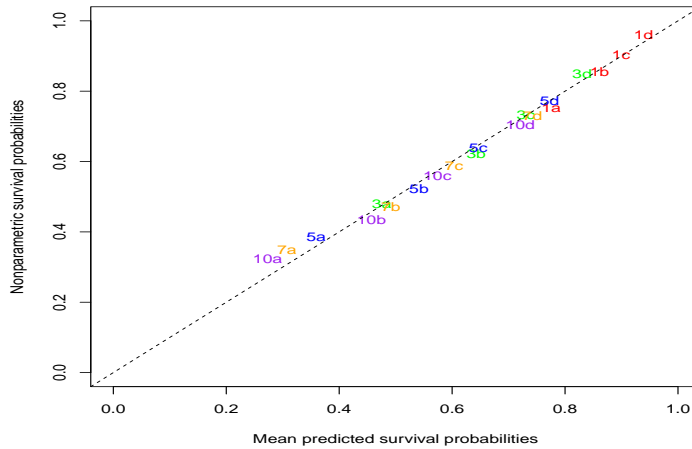


Figure 3.5 Calibration plot

3.4.2. AUC(t) 결과

적합된 비례위험모형의 판별적 정확성을 검증하기 위해 1, 3, 5, 7, 10년에서 AUC(t)를 계산하였다. Table 3.4에서 AUC(t) 값들이 대략 0.7이므로, 이는 적합된 Cox 비례위험모형의 판별적 정확성이 좋음을 의미한다.

Table 3.4 AUC(t) over 1, 3, 5, 7, and 10 years

Year	AUC(t)
1	0.708
3	0.689
5	0.677
7	0.672
10	0.668

4. 결론

미국 국립암연구소의 SEER 프로그램에서 제공하는 우측 대장암 3기 자료에 Cox 비례위험모형을 적합하여 생존율을 예측하기 위한 모형을 구하였고, 모형에 사용된 공변량들이 비례위험 가정을 만족함을 확인하였다. 적합된 모형의 타당성을 검증하기 위해 10-fold 교차 검증을 이용하여 생존함수를 추정하였고 이 추정치들을 이용하여 calibration 도표와 시간에 의존하는 AUC(t)를 계산하였다. 그 결과 적합된 Cox 비례위험모형이 좋은 검정력과 판별력을 가졌음을 확인하였다.

Cox 비례위험모형 적합 결과, 나이가 많은 환자일수록 사망 위험률이 높았으며, 남성보다는 여성의 사망 위험률이 낮았다. 기혼자보다는 미혼자의 사망 위험률이 높았으며 이는 Gove (1973)의 연구결과와 일치한다. 진단연도가 최근일수록 사망 위험률이 낮았으며 이는 암 치료법의 발전에 따른 결과일 것으로 판단된다. 글리슨 점수가 높을수록 우측 대장암 3기 환자의 사망 위험률이 높았다. 세분화된 병기는 다른 공변량들의 회귀계수 추정치보다 월등히 크므로 세분화된 병기가 사망률을 예측함에 있어서 가장 큰 영향을 끼침을 알 수 있었다.

본 논문에서 우측 대장암 3기 자료의 생존모형을 찾고, 그 모형의 적합성을 검토하기 위해 사용되어진 분석 방법들은 다른 암 자료의 분석에도 사용될 수 있을 것이다. Lee 등 (2015)은 생존분석을 이용하여 직장암 연구에 대한 결과를 보여주었다. 본 연구에서는 우측 대장암 자료의 생존분석 결과만 보여주

었으나, 좌측 대장암과 직장암도 분석에 포함시킨다면 암이 발생하는 위치 (site)를 Cox 비례위험모형에서 공변량으로 고려하여 암의 발생 위치가 암 사망률에 미치는 영향을 조사해 볼 수 있다. 또한 항암 치료, 방사선치료에 대한 정보나 식이요법과 같은 정보가 제공된다면 이들 또한 공변량으로 고려하여 이 요인들이 암 사망률에 미치는 영향을 조사해 보는 것도 필요할 것이다. 이는 본 연구의 향후 연구과제가 될 것이다. 본 연구에서는 분석에 사용된 공변량들이 비례위험 가정을 만족하여 추가적인 분석이 이루어지지 않았지만, 비례위험 가정을 만족하지 않는 공변량이 있을 경우 층화된 비례위험모형을 적용하거나 Cox 비례위험모형에 시간 가변 공변량 (time-dependent covariate)을 포함시키는 것을 고려할 수 있다 (Kim과 Lee, 2013). 이는 본 연구의 향후 연구과제가 될 것이다. 본 연구에서는 우측 대장암 3기 환자가 암에 의해 사망하는 경우만 고려하였다. 그러나 고령의 암 환자의 경우 암과 관련 없는 다른 원인에 의해서도 사망할 수 있다. 이와 같이 암 환자가 암 뿐만 아니라 다른 원인에 의해서도 사망할 수 있는 경우 경쟁위험모형 (competing risks model)을 적용하여 (Prentice 등, 1978; Fine과 Gray, 1999; Doh 등, 2015) 암에 의한 사망과 다른 원인에 의한 사망에 영향을 미치는 요인에 대해 조사할 수 있으며, 이는 본 연구의 향후 연구 과제가 될 것이다.

References

- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89-99.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B (Methodological)*, **34**, 187-220.
- Doh, G., Kim, S. and Kim, Y. J. (2015). Statistical analysis of economic activity state of workers with industrial injuries using a competing risk model. *Journal of the Korean Data & Information Science Societies*, **26**, 1271-1281.
- Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for subdistribution of a competing risk. *Journal of the American Statistical Associations*, **94**, 496-509.
- Gleason, D. F. (1992). Histologic grading of prostate cancer: A perspective. *Human Pathology*, **23**, 273-279.
- Gove, W. (1973). Sex, marital status, and mortality. *American Journal of Sociology*, **79**, 45-67.
- Grambsch, P. and Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.
- Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, **61**, 92-105.
- Iacopetta, B. (2002). Are there two sides to colorectal cancer? *International Journal of Cancer*, **101**, 49-60.
- Kaplan, E. L. and Meier P. (1958). Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.
- Kim, Y. and Lee, H. (2013). Estimation of lapse rate of variable annuities by using Cox proportional hazard model. *Journal of the Korean Data & Information Science Society*, **24**, 723-736.
- Koh, S. and Kim, J. (2010). The reasons for the increased incidence of colorectal cancer in Korea. *The Korean Journal of Medicine*, **79**, 97-103.
- Lee, M., Cronin, K. A., Gail, M. H. and Feuer, E. J. (2012). Predicting the absolute risk of dying from colorectal cancer and from other causes using population-based cancer registry data. *Statistics in Medicine*, **31**, 489-500.
- Lee, S., Shim, B. and Kim, J. (2015). Estimation of hazard function and hazard change-point for the rectal cancer data. *Journal of the Korean Data & Information Science Society*, **26**, 1225-1238.
- Miller, M. E., Langefeld C. D., Tierney W. M., Hui S. L. and McDonald C. J. (1993). Validation of probabilistic predictions. *Medicine Decision Making*, **13**, 49-58.
- Ministry of Health and Welfare. (2014). Cancer registration statistics, <http://ncc.re.kr/cancerStatsList.ncc?>.
- National Cancer Information Center. (2014). Cancer information service: Colon cancer, <http://www.cancer.go.kr/mbs/cancer/>.
- National Cancer Institute. (2016). Surveillance, Epidemiology, and End Results (SEER) Program, <http://seer.cancer.gov/>.

- Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34**, 541-554.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.
- Vergouwe, Y., Steyerberg, E. W., Eukemans, M. J. and Habbema, J. D. (2002). Validity of prognostic models: When is a model clinically useful? *Seminars in Urologic Oncology*, **20**, 96-107.
- World Health Organization. (2016). International classification of diseases, <http://www.who.int/classifications/icd/en/>.

Analysis of stage III proximal colon cancer using the Cox proportional hazards model[†]

Taeseob Lee¹ · Minjung Lee²

^{1,2}Department of Statistics, Kangwon National University

Received 22 February 2017, revised 20 March 2017, accepted 22 March 2017

Abstract

In this paper, we conducted survival analyses by fitting the Cox proportional hazards model to stage III proximal colon cancer data obtained from the Surveillance, Epidemiology, and End Results program of the National Cancer Institute. We investigated the effect of covariates on the hazard function for death from proximal colon cancer in stage III with surgery performed and estimated the survival probability for a patient with specific covariates. We showed that the proportional hazards assumption is satisfied for covariates that were used to analyses, using a test based on the Schoenfeld residuals and plots of the Schoenfeld residuals and $\log[-\log\{\widehat{S}(t)\}]$. We evaluated the model calibration and discriminatory accuracy by calibration plot and time-dependent area under the ROC curve, which were calculated using 10-fold cross validation.

Keywords: Calibration plot, Cox proportional hazards model, proximal colon cancer, ROC curve, survival analysis.

[†] This work was supported by Basic Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future planning (2016R1C1B1010294).

¹ Graduate student, Department of Statistics, Kangwon National University, Chuncheon 24341, Korea

² Corresponding author: Assistant professor, Department of Statistics, Kangwon National University, Chuncheon 24341, Korea. E-mail: mlee@kangwon.ac.kr