The study on survivability of Korean-American Lung cancer patients in SEER data with Kaplan Meier Method

Daegon Cha

Department of Industrial Engineering, the University of Ajou, Suwon

Abstract

According to 2012 WHO (World Health Organization) Health Data, not only lung cancer ranks at the most common cancer in the world, but the cancer also is the deadliest one among all type of cancer in the world. Therefore, it is imperative to analyze causing factors of lung cancer and study the best medical treatment against threat of lung cancer. In order to provide fundamental research for lung cancer, I focus on survivability of Korean-American by utilizing SEER (Surveillance Epidemiology End Result program) data. Also, results of this study will be expanded to further researches about defining differences of attributes impacting on survivability between KA(Korean-American) lung cancer patients and those of Korean. In this research, I studied survivability of KA with lung cancer in United States and then figuring out how three factors such as age, grade and types of lung cancer affect survivability of lung cancer patient. I selected 601 cases of KA with lung cancer from over 4.7 million datasets in SEER data and used Kaplan Meier method to examine 601 patient's survivability.

Keywords: Kaplan Meier, Lung cancer, SEER

1. Introduction

폐암은 전세계적으로 가장 흔하게 나타나는 암 질환 중의 하나로, 남자와 여자 모두에게서 암으로 인한 사망의 제 1 원인으로 알려져 있다. [1] 폐암의 가장 큰 원인은 흡연이며, 폐암 위험은 흡연량과 흡연기간에 따라 달라진다. 그 외에는 간접흡연, 직업성 이나 환경적으로 라돈, 석면, 크롬, 비소, 카드뮴에 대한 노출과 더불어 방사선, 대기오염 또는 실내에서 발생하는 연기가 원인이 될 수 있다. [2]

WHO 의 조사에 따르면, 미국에서 폐암의 5 년 생존율은 16%이고, 전이가 일어나지 않은 채 진단된 원발성 폐암의 생존율은 53%이다 [3].

또한, 우리나라의 경우 2013 년 통계청이 발표한 사망원인 통계에 따르면, 인구 10 만명당 암에 의한 사망률은 149 명으로 그 중, 폐암 (34 명), 간암 (22.6 명), 위암 (18.2)명 이였으며, 폐암 사망률은 2012 년 대비 전년대비 2.7% 증가 하였다. [4]

폐암은 조기진단이 어렵고, 종양이 발견된다고 해도 이미 많이 진행되어 있는 경우가 많기 때문에, 주요 선진국에서도 생존율이 높지 않는 편이다.

국내 폐암 생존율에 대한 연구로써, 각 의료기간에 보관되어있는 폐암환자 데이터를 기반한 폐 절제수술 후의 환자들의 생존율 분석과 병기에 따른 생존율 분석에 국한되어 있는 경우가 많았다.

이러한 연구방식의 한계점을 극복하고자 본 연구의 목적은 여러 의료기관에서 수집한 포괄적인 한국인 폐암환자의 데이터를 기반으로 생존기간을 분석하는 데에 있다.

하지만 국내의 의료정보보호법과 공공의료 데이터 획득이 용이하지 않는 환경을 감안하여, 미국 국립 암 센터 (National Cancer Institute) 에서 공개적으로 제공하는 SEER(Surveillance Epidemiology and End Results) 데이터를 활용하여 연구를 진행한다.

본 연구에서는 1971 년부터 현재까지 축적된 약 470 만건의 SEER 데이터 중에서, 암의 병기(Stage)가 명확히 확인된 한국인 교포 폐암환자 601 명의 데이터를 추출하고, 각 요인에 맞는 데이터 전처리 작업과 생존기간 분석기법인 Kaplan Meier method 를 통해 나이, 병기(Stage) 그리고

폐암의 종류가 환자의 생존기간에 얼마나 요인을 미치는지 연구하고자 한다.

2. Methods

2.1 데이터 선정 및 요인선정.

본 연구는 약 470 만건의 SEER 암환자 데이터 중에서 Stage1 에서 Stage4 까지의 병기가 명확히 확인된 한국인 폐암환자 601 명을 대상으로 진행하였다.

요인으로 폐암환자의 나이, 병기(Stage) , 폐암의 세부 종류들로 총 3 가지 요인을 바탕으로 생존율을 분석하였다.

특히, 폐암의 세부종류의 분류는 전세계 폐암발생 중 80%~85%를 차지하고 있다고 보고되는 비 소세포 성 폐암의 3 가지 종류 즉, 선암, 편평 세포암, 대 세포암으로 폐암을 세분화하여 각 종류에 대해 생존율을 분석하였다.

2.1.1 선정된 샘플의 특성분석

Table 1. Characteristics of 601 Cases

Sex			**Percentage**
	Male	311	51.7%
	Female	290	48.3%
	Total	**601**	
Grade			
	Stage1	50	8.3%
	Stage2	164	27.3%
	Stage3	309	51.4%
	Stage4	78	13.0%
	Total	**601**	
Age Group			
	20~39	7	1.2%
	40~59	150	25.0%
	60~79	366	60.9%
	80~99	78	13.0%
	Total	**601**	
Cell Type of Lung Cancer (*ICD #)			
**	8140	207	52.9%
***	8070	158	40.4%
****	8012	26	6.6%
	Total	**390**	

*국제질병분류코드(International Classification of Disease)
**선암 (Adenocarcinoma)
***편평 세포암 (Squamous Cell Carcinoma of Lung)
****대 세포암 (Large-Cell Carcinoma of Lung)

2.2 데이터 분석기법

본 연구에서 사용하는 Kaplan Meier 란, 생존기간을 분석하여 생존곡선을 추정하는 통계기법으로 치료방법, 예후인자 등이 생존에 미치는 효과를 추정하는데 사용된다. Kaplan Meier 는 생존여부와 생존기간, 그리고 생존율에 영향을 끼칠 것이라고 생각되는 요인 하나 즉, 3 가지의 데이터를 기반으로 생존율을 나타낸다.

SEER 데이터에서 나타난 생존여부는 환자의 암이 환자의 죽음에 직접적인 영향을 끼친 경우에만 '사망(Death)'로 기록하였다. 또한 생존기간은 0 부터 9998 개월까지의 생존한 Month 를 기록을 하며, 환자의 생존기간을 알 수 없는 경우에는 9999 로 기록되어 있다. 따라서 본 연구에는 9999 로 기록된 케이스를 제외한 나머지 케이스로 환자의 생존기간을 나타내었다. 또한 본 연구에서는 P value < 0.05 일 때 통계학적으로 유의 하다고 판단하였다.

2.3 각 요인 별 생존율 결과 화면구성

요인 별 생존율 도출화면은 총 3 가지로 구성되어 있다. 첫 번째는 각 요인의 총 케이스 수 및 사망 (Event) 가 발생 수를 정리한 케이스 요약이다. 두 번째는 생존시간에 대한 평균 및 중위 수를 나타내는 표이며, 마지막으로 Kaplan Meier 기법으로 나타난 결과가 계단식 생존함수로 나타난다. Results 부분에서 각 요인 별 결과값을 나타낼 예정이다.

3. Results

3.1 나이그룹별(Age Group) 생존율 분석 및 결과.

SEER 데이터에서 기본적으로 제공하는 나이분류는 총 18 가지로 분류되어있다. 이는 0 세부터 84 까지 5 살단위로 끊어 총 17 가지의 그룹으로 묶여있고, 마지막 18 번째 그룹은 85 세 이상으로 되어있다. 하지만 본 연구의 대상이 되는 601 명의 한국 교포의 폐암환자 특성을 고려하였을 때, 기존 SEER 에서 제공하는 나이그룹은 과도하게 많기 때문에, 생존함수 도표의 가독성 향상을 위해 20 살 단위로 축소하여 총 4 개의 그룹으로 재코딩 하였다. 밑의 Table 2 안의 왼쪽 표는 기존 SEER

데이터에서 기초 나이별 구분이고, 오른쪽 표는 20 살 단위로 다시 분류한 나이별 그룹이다.

Table 2. Code Description for Age group

*SEER 나이 구분

코드	구분
00	Age 00
01	Age 01-04
02	Age 05-09
03	Age 10-14
04	Age 15-19
05	Age 20-24
06	Age 25-29
07	Age 30-34
08	Age 35-39
09	Age 40-44
10	Age 45-49
11	Age 50-54
12	Age 55-59
13	Age 60-64
14	Age 65-69
15	Age 70-74
16	Age 75-79
17	Age 80-84
18	Age 85+
99	Unknown

** 변경 후 나이 구분

코드	구분
01	Age 20-39
02	Age 40-59
03	Age 60-79
04	Age 80-99

환자의 나이그룹 분류작업을 완료한 후 SPSS 의 Kaplan Meier 생존율 분석을 실시하였다. 2 개의 결측 값을 제외한 나머지 599 건이 활용되었다.

Table 3. Case Description for Age group

Age group	# of Case	사건(Death)	퍼센트
1 (20-39)	7	1	85.7%
2 (40-59)	150	9	94.0%
3 (60-79)	365	37	89.9%
4 (80-99)	77	9	88.3%
전 체	599	56	90.7%

60 세부터 79 세까지의 환자가 365 명으로 가장 많았으며, 전체적으로 나이에 상관없는 높은 생존율을 보여주었다. 하지만 Table4 에 나타난 것 같이 평균 생존 값은 나이와 비례해 낮아지는 것을 볼 수 있다. 이에 관련해 나이 그룹 별 생존율에 대한 자세한 내용은 아래의 Table 4 의 나이 그룹별 평균 생존시간 및 중위 수 (Median) 에서 확인할 수 있다.

Table 4. Average Survivability for Age group

Age group	평균생존 값	중위 수
1 (20-39)	96	96
2 (40-59)	353.083	351
3 (60-79)	171.383	167
4 (80-99)	82.350	83
전 체	232.646	278.419

위의 표를 보면, 20 에서 39 사이의 나이 그룹은 평균 생존 값이 상대적으로 높아 보이진 않지만, 적은 표본 수(7 명)와 사망자 (1 명), 90%에 가까운 중도절단 (추적불가)을 고려해 봤을 때, 이 평균생존 값이 해당 연령대의 평균생존 값을 대표한다고 보긴 힘들다. 그룹 1 을 제외한 나머지 나이 그룹별 평균생존 값은 나이별 그룹이 높아짐에 따라 전 그룹 대비 약 50%씩 생존 값이 감소하는 경향을 볼 수 있었다.

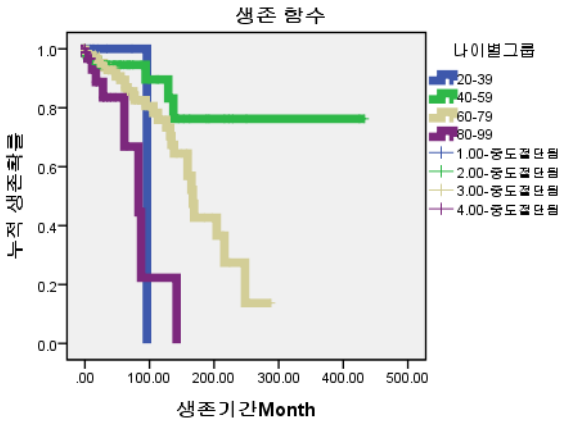

Figure 1. Survivability Function for Age Group

생존함수에 나타난 X 축은 생존기간이며, Y 축은 생존기간에 따른 누적생존율을 나타낸다. 나이별 그룹 1 (20 세~39 세)은 100 개월이 넘어가면서 급격한 절단면이 나타나 해당 나이별 그룹 샘플 7 명 중 1 명의 사망자를 제외한 6 명에 대한 추적이 끝난 것으로 추측해 볼 수 있다. 나이별 그룹 2(40 세~59 세)는 그룹 1 을 제외한 나머지 그룹 중 중위 수 29 년에 가까운 월등한 생존율을 보여주었으며, 그룹 3 과 4 는 각각

대략 13 년과 7 년으로 급격히 낮아지는 것을 볼 수 있었다.

3.2 폐암 Grade 별 생존율 분석 및 결과.

SEER 데이터에서 초기 암 병기 분류는 총 9가지로 나뉘어져 있다. 1번부터 4번까지 암의 진행상태인 1기에서 4기를 나타내며, 5번부터 8번까지는 특수한 형태를 가진 암을 분류해 놓은 코드이다. 마지막으로 9번은 진행상태를 정확히 식별하기가 불분명한 암을 분류해놓은 코드이다.

서론에서 논의 했듯이, 본 연구는 암의 진행상태가 확실히 구별이 되는 1기에서 4기에 해당하는 한국인 폐암환자 601명을 대상으로 진행 하였다.

Table 5. SEER Code Description for Grade

Code	Description
1	Grade 1;well differentiated
2	Grade 2; moderately differentiated
3	Grade 3; poorly differentiated
4	Grade 4; undifferentiated
5	T-Cell
6	B-Cell
7	Null cell
8	N K cell
9	Stage between Grade 2 and 3

SEER 데이터 중 전체 한국인 암환자 10647명 중에서 폐암의 진행상태가 명확한 환자 601명을 선정하여 Kaplan Meier 분석을 진행 하였다.

Table 6. Case Description for Grade

Grade	# of Case	사건 (Death)	퍼센트
Stage 1	50	7	86.0%
Stage 2	164	16	90.2%
Stage 3	307	23	92.5%
Stage 4	78	10	87.2%
전 체	599	56	90.7%

3기의 폐암이 307건으로 1기에서 4기로 판명된 한국인 폐암환자 중 가장 많은 케이스로 나타났고, 전체 599명중 56명이 사망 하였다. 폐암 기 별 평균 생존시간 및 중위 수는 Table 7에서 확인 할 수 있다

Table 7. Average Survivability for Grade

Grade	평균생존 값	중위 수
Stage 1	256.96	169
Stage 2	251.05	216
Stage 3	181.185	167
Stage 4	134.28	126
전 체	232.64	169

위의 표를 참조해보면, Stage가 진행될수록 평균 생존 값이 낮아지는 것을 확인할 수가 있다. 특히 Stage1과 2사이에는 큰 생존 값의 차이를 보이진 않지만, Stage2와 3, Stage 3과 4사이의 평균 생존 값의 차가 현격히 커지는 것을 알 수 있었다.

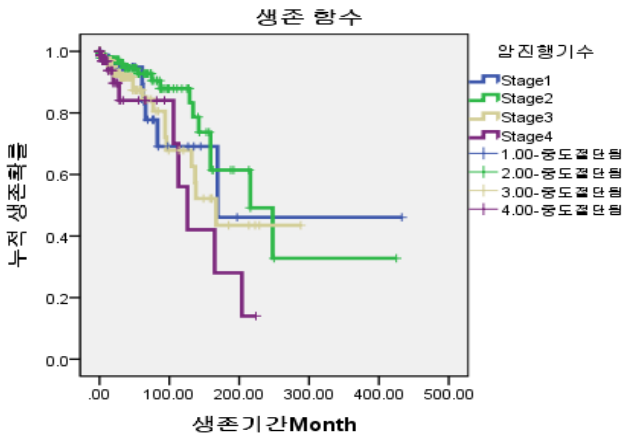

Figure 2. Survivability Function for Grade

생존함수를 보면 알 수 있듯이, 병기의 진행이 낮을수록 높은 생존 값을 보였다. Stage1 의 평균생존 값과 Stage4 의 평균 생존 값을 비교해 보면 대략 누적생존확률이 40%이상 차이 나는 것을 알 수 있다. 따라서 Stage 의 진행여부가 폐암환자의 생존율에 유의미한 영향을 끼친다고 볼 수 있다.

3.3 폐암의 세부종류 별 생존율 분석 및 결과

서론에서 논의했던 것처럼 전세계 폐암발생 중 80%~85%를 차지하고 있다고 보고되는 비소세포 성 폐암의 3 가지 종류 즉, 선암, 편평세포암, 대세포암 으로 세분화하여 각 종류에 대해 생존율을 분석 하고자 한다.

이를 위해, 전체 SEER 데이터에서 한국인 (Race code: 8) 이면서 폐암환자 (Cancer Code:

C340~C349)인 601 명을 다시 비소세포성 폐암의 하위그룹인 선암 (ICD#:8140), 편평세포암 (ICD#:8070), 대세포암 (ICD#:8012)으로 재 분류하여 각각 고유 코드를 부여하였다.

Table 8. Case Description for subtypes of non-small cell lung cancer

Code and Type	# of Case	사건(Death)	퍼센트
1 (선암)	207	17	91.8%
2 (편평세포암)	157	15	90.4%
3 (대세포암)	26	4	84.6%
전 체	390	36	90.8%

전체 한국인 폐암환자 601명중 세부 비 소세포 성 폐암을 가진 환자로 분류한 결과, 이 분류에 해당되는 환자는 390명으로, 가장 많이 보고된 암은 선암 (207명) 이였다.

Table 9. Average Survivability for subtype of non-small cell lung cancer

Code	평균생존 값	중위 수
1 type	287.5	284
2 type	180.46	167
3 type	129.88	126
전 체	249.98	192

Table 7을 보면 알 수 있듯이, 제 1 타입으로 코딩 한 선암 (Adenocarcinoma) 환자는 평균생존 값이 287.5개월로 하위 세 개의 비 소세포성 폐암 중 타 암에 비해 월등히 높은 생존 값을 가지고 있었다. 그 뒤로 제 2 타입인 편평세포암 (Squamous Cell Carcinoma)와 제 3 타입인 대세포암(Large Cell Carcinoma)은 각각 180개월, 129개월의 생존 값을 보여주었다.

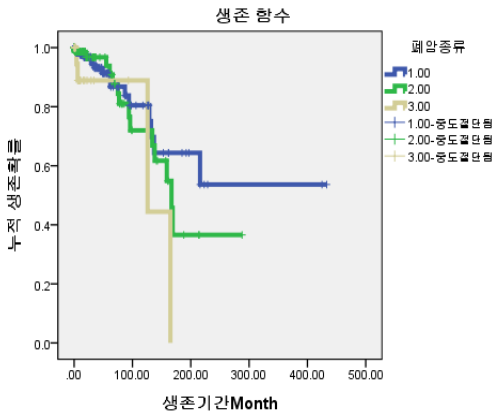

Figure 3. Survivability Function for subtypes of non-small cell lung cancer

Figure 3 에서 알 수 있듯이 제 1 타입인 선암의 경우 평균생존 값이 다른 종류의 비 소세포성 폐암보다 월등히 긴것으로 나타났다. 따라서, 전체 폐암 환자 중 80%에서 85%를 차지하는 비 소세포 성 폐암 중 선암 일 경우 환자의 높은 생존 값에 지대한 영향을 미칠 것으로 파악 되었다

4. Discussion

본 연구에선 SEER 데이터 내 한국인 폐암 환자의 생존기간에 영향을 미치는 요인으로 Age, Grade 그리고 비소세포성 암의 종류인 선암, 편평세포암 그리고 대세포 암을 요인으로 지정하여 Kaplan Meier Method 로 생존분석을 진행하였다.

Age에 따른 생존기간은 나이별 그룹이 올라가 수록 전 그룹 대비 50% 정도의 생존기간이 하락하는 양상을 보였고 이는 Age가 폐암환자의 생존기간에 유의한 영향을 미친다고 볼 수 있다.

Grade 요인 에서는 Stage1과 Stage2의 폐암환자 간 생존기간 차이는 5개월 차이로 유의한 변화가 없다가, Stage2 에서 Stage3으로 넘어가는 단계에서 70개월 차이를 보이는 급격한 생존기간의 하락이 눈에 띄었다. 이는 생존기간을 높이기 위해선 폐암의 심각한 진행을 발견하기 위한 조기검진이 필요하다는 것을 시사해 준다.

마지막으로, 비소세포성 폐암 세부 종류별 생존기간 분석 결과, 선암(Adenocarcinoma)의 경우 평균 생존기간이 287.5 개월로 다른 종류의 비소세포성 폐암보다 대략 150개월이 더 긴 것을 알 수 있었다.

이번 연구로 본 연구에서 선택한 3개의 변수 (Age, Grade and subtype of non -small cell lung cancer) 모두가 SEER 데이터 내 한국인 폐암환자의 생존기간에 영향을 미치는 것으로 결론 내었다. 또한 예상과는 달리 SEER 데이터 내의 한국 교포 폐암환자의 생존기간이 예상했던 것 보다 높았던 것이 인상적 이였다. 향후, 이 연구를 바탕으로 국립 암 센터의 폐암환자 Cohort DB에서 추출한 결과와의 비교연구를 진행하여, 한국 교포와 한국인과의 생존기간에 미치는 요인의 차이점을 도출하는 연구를 진행 할 예정이다.

5. References

[1] International Agency for Research on Cancer.
GLOBOCAN 2012: Estimated cancer incidence, mortality
and prevalence worldwide in 2012.
(Retrieved from: http://globocan.iarc.fr 9-12-2015)
[2] Global Cancer Facts & Figure 2nd Edition 15 page
[3] Global Cancer Facts & Figure 2nd Edition 18 page
[4] 2013 년 사망원인 통계 (통계청 2013) 10 page.