

# Analyzing Brain Cancer Survivability Using SEER(Surveillance, Epidemiology, and End Results Program) Data with method of Kaplan Meier

Seungmin Lee<sup>1</sup>, Beom Park

Department of Industrial Engineering, the University of Ajou, Suwon, 16497

## ABSTRACT

In Korea, there were 75,334 deaths caused by cancer and more than 1,000 cases were caused from brain cancer in 2013. An increasing incidence of brain cancer has been reported for the last three decades and the survival rate of brain cancer for 2 years is only about 20percent. Therefore, understanding the causes of brain cancer and suggesting the best medical treatment for individual have become one of major studies at this time. This paper will mainly focus on understanding the significant factors of causing brain cancer and analyze how much they influence on survival rate. In order to understand the significant factors, the analysis of patient's medical record has been performed. SEER data, provided by national cancer institute in America, was used as the main data for the study. The data contains 62,029 cases of brain cancer and 134 variables of patients' information; sex, race, treatment history etc. For the analysis method, Kaplan Meier was mainly used to make a division between correlated factors and uncorrelated factors. The analysis have performed with 4 factors that have widely known to be the most influential to survival rate of cancer; age, sex, surgery performed and grade.

Keywords: Brain cancer, Survivability, SEER Data.

## 1. Introduction

국내 대다수의 암 관련 연구는 주요 5대암을 기준으로 이루어지고 있다. 주요 5대암으로는 갑상선, 위, 대장, 폐, 유방암으로, 1993년부터 2014년까지 생존율이 적게는 10%에서 많게는 30%정도 증감되었다.[1] 이처럼 주요 5대암에 대한 생존율이 크게 증가한대에는 해당 암에 대한 다양한 요인분석과 연구를 통해, 맞춤치료 및 조기진단이 가능하게 되었기 때문이다.[2]

뇌종양은 국내 주요 5대암에 비해 발병률이 전체 암의 2%로 낮은 편이지만, 5년 사망률이 90%때로 굉장히 저조한 편이며, 치료에 있어 뇌라는 특수성 때문에 많은 어려움을 겪고 있다.[3] 또한, 해외에서는 뇌종양에 대한 연구가 활발히 진행되고 있는 반면, 국내에서의 뇌종양 관련 연구는 부족할 실정이다. 하지만 최근 국내에도 뇌은행이 설립되고, 뇌종양의 조기진단 및 맞춤치료를 위한 연구가 진행되면서 중요성이 부각되고 있다.

본 연구는 국내의 뇌종양 연구에 기여하고자,

해외표본 암 데이터인 SEER데이터를 활용하여 뇌종양 생존율에 영향을 미치는 요인들을 분석하였다. 해외데이터인 SEER데이터를 활용하는 이유는 국내 뇌종양 환자들에 대한 공공데이터가 부족하여 충분한 데이터 확보가 어렵고, 의료정보 규제에 따른 병원들의 병리정보 활용도 힘들기 때문이다. SEER데이터의 경우 미국의 20세미만 청소년들의 암 사망률 1위가 뇌종양인 만큼, 뇌종양에 대한 데이터가 많고, 환자정보와 암의 특성을 요인 별로 정리하여 분석하는데 용이하다.[4]

본 연구 방향은 뇌종양의 대한 생존율을 요인 별로 분석하기 위하여 대략 460만건의 SEER데이터 중 뇌종양의 데이터인 62,029건의 데이터를 추출하고, Kaplan Meier분석기법을 활용해 요인 별로 얼마나 생존율에 많은 영향을 미치는지 분석하였다. 선정된 요인으로는 전문가들의 보편적 의견을 반영해 Grade, Age, Sex, Surgery 이 4가지로 정하였고, 각 요인마다 데이터 정제작업을 거친 후 생존율을 도출하였다.[5~7]

## 2. Method

### 2.1 데이터 분석기법 및 요인선정

Kaplan Meier 생존분석 기법은 사망이 발생함에 따라 누적되는 형태의 계단형태 생존곡선을 나타내는 기법이다.[6] 따라서, 본 연구는 62,029건의 뇌종양 데이터를 요인 별로 정리하고, Kaplan Meier를 활용하여 평균 생존율을 도출 및 가장 생존율에 큰 영향을 미치는 요인을 정리할 것이다. 비교요인들 선정은 전문의들이 보편적으로 핵심요인이라 지목하는 Grade, Age, Sex, Surgery 이 4가지 요인이다.

### 2.2 Kaplan Meier 분석기법 평균 생존율 도출

SPSS 통계툴의 Kaplan Meier 분석기법을 활용하기 위하여 데이터 정제작업을 진행하였다. Kaplan Meier는 생존여부, 생존기간 그리고 요인 이 3가지를 기반으로 생존율을 도출한다. 생존여부의 경우 환자가 해당 암 때문에 사망했을 경우만을 Death하였다. 생존기간의 경우 신뢰성 있는 데이터를 얻기 위한 핵심 요소이기 때문에 SEER 데이터에서 권장하는 방법으로 코딩작업을 실시하였다.

#### 생존기간 수식

$$\text{Survival months} = \text{Floor}((\text{end point} - \text{date of diagnosis}) / \text{days in a month})$$

생존기간 수식 및 정제작업은 4가지 요인분석에 모두 동일시 적용되었다. 그 외의 데이터 정제작업은 각 요인마다 필요한 부분에 진행되었으며, 자세한 사항은 요인별 결과에 언급하였다.

### 2.3 SEER 데이터 생존율 도출 결과 화면

생존율 도출 결과화면은 총 3가지로 나누어진다. 첫 번째는 각 요인의 변수의 총 케이스 수 및 사건(Death)이 발생한 수를 정리한 케이스 처리 요약이다. 두 번째는 생존시간에 대한 평균 및 중위수를 나타내는 표이며, 마지막으로 Kaplan Meier 기법을 활용한 생존함수가 결과화면에 포함될 것이다.

## 3. Results

### 3.1 Grade별 평균 생존율 및 결과

SEER 데이터에서 제공하는 Grade는 종류별 총

9가지이다. 1번부터 4번의 경우 암의 진행상태인 1~4기를 나타내며, 5번부터 8번까지의 경우는 암의 모양에 따라 분류하였다. 마지막 9번은 2기와 3기 사이에 분류가 애매한 경우이다. 따라서, 본 연구에서는 생존율에 직접적인 영향을 미치는 암의 크기(1기~4기)와 2기와 3기사이인 9번만을 포함하였다. 표1은 SEER 데이터의 Grade별 생존율을 도출하기 위해 분류별로 코딩한 내용이다.

Table 1. 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

코딩작업을 완료한 후 Kaplan Meier 생존율 분석을 하였다. 코딩작업에서 정제된 일부 데이터를 제외한 60,654건이 활용되었다. Table 2는 Grade별 Case number와 사건 수(Death)를 보여준다.

Table 2. Case Description for grade

Grade	합계 N	사건 수	퍼센트
1	1541	448	70.9%
2	4474	2163	51.7%
3	4331	3361	22.4%
4	14171	10099	28.7%
9	36137	20779	42.5%
전체	60654	36850	39.2%

Grade 9이 36,137건으로 Brain Cancer에서 가장 많은 Case로 나타났고, 전체데이터 60,654건 중 사망률이 60퍼센트에 이르는 36,854건이 사망에 이르렀다. Grade별 생존율에 대한 자세한 사항은 아래 Table 3의 Grade별 평균 생존시간 및 중위수에서 확인할 수 있다.

Table 3. Mean Survivability for Grade

Grade	평균 생존 값	중위수 평균
1	312.518	161
2	223.717	138
3	65.805	11
4	66.124	12
9	144.19	17
전체	131.828	17

Table 3의 평균 추정값이 중위수의 평균 추정값보다 월등히 높은 이유는 생존환자들의 평균 생존기간이 300months를 넘어 전체적인 평균값을 상향시켰기 때문이다. 따라서 정확한 결과를 보기 위해선 중위수의 평균 생존값을 보아야 한다. 테이블에서 쉽게 볼 수 있듯이 Grade 2기를 넘어간 시점부터 생존기간이 급격하게 줄어드는 것을 볼 수 있다. Grade 2의 평균생존기간이 138month에 비해 Grade 3의 생존기간은 11month로 생존율에 Grade가 지대한 영향을 미치는 것을 알 수 있다. Table 3의 결과값을 생존곡선으로 나타낸 것이 Figure 1이며, Grade가 생존율에 미치는 영향을 쉽게 볼 수 있다.

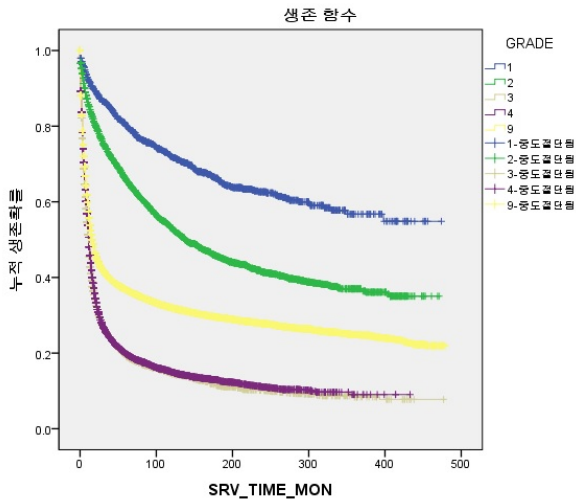


Figure1. Survivability Graph for grade

X축은 생존기간이며, Y축은 생존기간에 따른 누적생존율을 나타낸다. Grade 1과 Grade 4를 비교하면 60개월을 기준으로 대략 70퍼센트의 생존율 차이를 볼 수 있다. 또한 Grade 2기를 넘어가는 시점에서는 생존율에 큰 변화가 없음을 알 수 있다.

### 3.2 Age별 평균생존율 및 결과

Age의 경우 0부터 130살까지 측정이 되었으며, 정확한 나이가 측정되지 않은 값에는 999값이 주어졌다. 따라서, 999값이 주어진 데이터는 코딩작업을 거쳐 정제하였으며, 총 60,656건의 데이터를 확보할 수 있었다. 나이는 총 5가지 값으로 분류하였다. 0~20세, 21~40세, 41~60세, 61~80세 그리고 81세 이상이다. 아래 Table 4에서 연령별 Case number와 사건수(Death)를 확인할 수 있다.

Table 4. Case Description for age

Age	합계 N	사건 수	퍼센트
0~20	8281	2610	68.5%
21~40	9761	4584	53.0%
41~60	17846	12089	32.3%
61~80	20761	14974	27.9%
81~	4007	2593	35.3%
전체	60656	36850	39.2%

61세부터 80세의 환자가 20,761명으로 가장 많았으며, 전체적인 생존율은 연령이 높을수록 낮아진 것을 볼 수 있다. 전체적인 연령별 사망률은 41이후로 비슷해 보이지만, Figure 5의 평균생존을 보면 확연한 차이를 볼 수 있다.

Table 6. Mean Survivability for age

Age	평균생존 값	중위수 평균
0~20	308.587	146
21~40	205.904	116
41~60	85.128	17
61~80	34.441	7
81~	25.781	3
전체	131.833	17

평균생존기간의 중위수 값을 보면 21~40세의 경우 116months로 10년정도의 생존기간을 나타낸다. 하지만 41세부터 60세는 17months, 61세부터 80세는 7months, 81세이상은 3months로 연령이 많을수록 생존기간이 기하급수적으로 내려가는 것을 확인할 수 있다. 생존함수인 Figure 2를 보면 연령별 생존기간의 차이를 쉽게 확인할 수 있다.

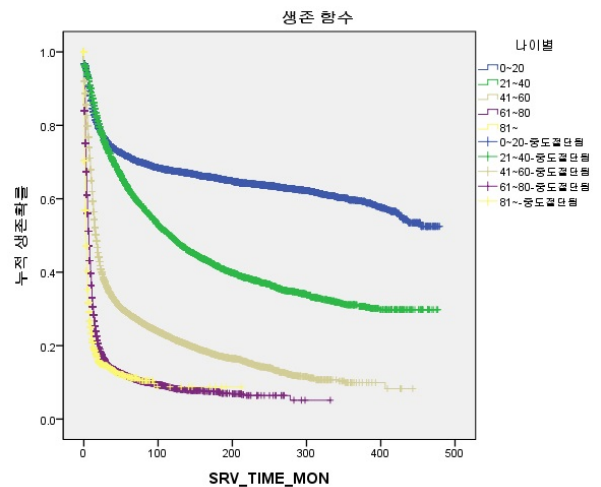


Figure 2. Survivability Graph for age

### 3.3 Sex별 평균생존율 및 결과

성별의 경우 1번이 male, 2번이 female로 분류되어있으며, 총 60,656의 케이스를 대상으로 생존율을 분석하였다. Table 7에서 성별 케이스 요약을 확인할 수 있다.

**Table 7. Case Description for sex**

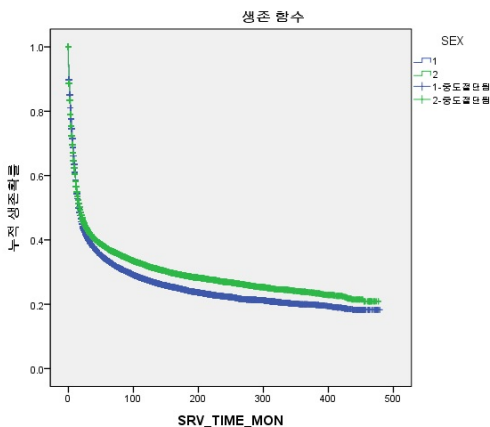
SEX	합계 N	사건 수	퍼센트
1	33671	20965	37.7%
2	26985	15885	41.1%
전체	60656	36850	39.2%

남성의 경우 33,671건으로 여성의 26,985건보다 대략 20퍼센트 정도 많았다. 전체적인 사건 수 (Deaths)는 36,850건으로 앞에 데이터들과 동일하다. (동일한 이유는 요인별로 사망자수를 분류한 것이기 때문에, 전체 사망자수에는 변함이 없다.) 사망자 비율을 보면 성별은 Brain Cancer에 큰 영향을 주지 않는 것을 확인할 수 있다. 이 부분은 Figure 8의 평균 생존시간을 보면 더욱 뚜렷하게 확인할 수 있다.

**Table 8. Mean Survivability for sex**

SEX	평균생존 값	중위수 평균
1	123.913	16
2	141.561	18
전체	131.833	17

평균 생존기간이 남성의 경우 16months, 여성의 경우 18months로 크게 차이가 없는 것을 볼 수 있다. 앞의 분석된 Grade와 Age에 비교해 생존율과 기간에 미치는 영향이 미비하다.



**Figure 3. Survivability graph for sex**

Figure 3의 결과도 앞의 결과들과 크게 다르지 않다. 남성과 여성의 생존곡선이 같은 추세로 내려가는 것을 확인할 수 있다.

### 3.4 Surgery여부별 평균생존율 및 결과

수술여부의 경우 총 8가지의 코딩작업이 필요하였다. 수술을 한 경우 0의 값을 주었고, 나머지 수술을 받지 않은 경우를 이유에 따라 1,2,5,6,7,8,9의 번호를 부여했다. 자세한 내용은 Table 8에 기재하였다.

**Table 8. Code Description for surgery**

Code	Description
0	Surgery performed
1	Surgery not recommended
2	Contraindicated due to other conditions
5	Patient died before recommended surgery
6	Unknown reason for no surgery
7	Patient or Patient's guardian refused
8	Recommended, unknown if done
9	Unknown if surgery performed

본 연구에서는 수술여부에 대한 생존율 도출이 목적이었기 때문에, 코드를 수술을 받은 환자(1번부여), 수술을 받지 않은 환자(2번부여)로 분류하였다. 수술을 받지 않은 환자의 데이터는 환자가 직접적으로 수술을 받을 수 있는 상황임에도 받지 않은 Case들만 포함시켰다. 따라서, 일부 적합하지 않은 데이터 코드들은 제외하였다. (제외된 Code: 2,5,6,9). Table 9의 Case Description에서 수술여부에 따른 케이스 수를 확인할 수 있다.

**Table 9. Case Description for surgery**

수술여부	합계 N	사건 수	퍼센트
1	40122	23594	41.2%
2	18572	11883	36.0%
전체	58694	35477	39.6%

적합하지 않은 데이터 케이스들을 제외한 58,694건 중 수술을 받은 환자 40,122건과 수술을 받지 않은 환자 18,572건의 데이터로 생존율을 분석하였다. Case Description에서 나타나듯이 수술을 받은 환자가 그렇지 않은 환자보다 월등한 생존율을 보여준다. Table 10의 평균생존기간 결과도 이와 같다.

**Table 10. Mean Survivability for surgery**

수술여부	평균생존 값	중위수 평균
1	147.447	22
2	100.254	9
전체	131.833	17

수술을 받은 환자의 경우 평균생존기간이 22months이지만, 수술을 받지 않은 환자의 경우 9months의 생존기간을 나타낸다. 이는 수술여부가 생존율과 생존기간에 큰 영향을 준다는 것을 알 수 있다.

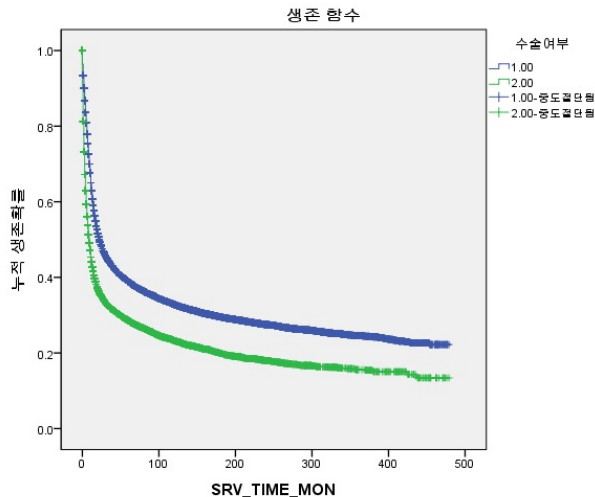


Figure 4. Survivability graph for surgery

Figure 4의 그래프를 보면 모든 구간에서 수술한 환자의 생존율이 높은 것을 확인할 수 있다. Grade와 Age와 같은 큰 차이를 보여주지는 않지만, 충분히 생존여부에 영향을 준다는 결론을 낼 수 있다.

#### 4. Conclusion

Brain Cancer의 생존기간에 영향을 미치는 요인 4가지를 분석한 결과 Grade, Age, Surgery 그리고 Sex 순으로 나타났다. Sex의 경우 생존기간의 차이가 2months로 사실상 영향이 없다고 봐도 무방하다. 한 가지 흥미로운 점은 Brain Cancer의 경우 Grade가 2기 이상을 넘어갈 경우 생존율이 급격히 낮아진다는 것이다. Brain Cancer의 생존율을 높이기 위해서는 조기 진단이 가장 중요한 요소라고 볼 수 있다. Age의 경우 60세 이상이 되는 시점에서 생존율과 생존기간이 비슷해 지는 것을 알 수 있다. 이는 60세 이상부터는 사실상 Age는 생존율에 영향을 크게 주지 않는다고 볼 수 있다. 마지막으로 수술여부는 Brain Cancer의 생존율에 평균 13months 정도의 영향을 준다는 결과가 나왔다. 수술을 받지 않은 환자의 평균 생존기간은 9months에 불과하지만, 수술을 받은 환자는

22months의 생존기간을 보였다. 본 연구는 SEER 데이터에 있는 134개의 요인 중 4개의 핵심 요인들을 바탕으로 생존율을 도출하였으며, 성별을 제외한 3개의 요인들이 생존율과 생존기간에 큰 영향을 준다는 결과를 얻었다. 추 후에는 이 연구결과를 바탕으로 각 요인들간의 상관관계를 분석하고, 이에 따른 생존기간을 도출하는 연구를 진행할 예정이다.

#### References

1. National cancer stat [www.cancer.net/cancer-types/brain-tumor](http://www.cancer.net/cancer-types/brain-tumor) (Retrieved September 5, 2015).
2. Brain tumor info [www.abta.org/about-us/news/brain-tumor-statistics](http://www.abta.org/about-us/news/brain-tumor-statistics) (Retrieved September 3, 2015)
3. Treatment of brain cancer [http://www.medicinenet.com/brain\\_cancer/article.htm](http://www.medicinenet.com/brain_cancer/article.htm) (Retrieved September 8, 2015)
4. SEER data <http://seer.cancer.gov/> (Retrieved 5, 2015)
5. Dursun Delen, Glenn Walker, Amit Kadam, *Predicting breast cancer survivability; a comparison of three data mining methods* 2004
6. National cancer Institute, <http://seer.cancer.gov/resources/> (Retrieved August 31, 2015)
7. Hyunjung Shin, Yonghyun Nam, *A coupling approach of a predictor and a descriptor for breast cancer prognosis*, 2013

#### Author listings

**Seung Min Lee:** darksween@msn.com

**Highest degree:** Bachelor, Department of Economics, Boston University

**Position title:** Researcher, Department of Industrial Engineering, Ajou University

**Areas of interest:** Ergonomics, Biomedical Informatics

**Beom Park:** ppark@ajou.ac.kr

**Highest degree:** PhD, Department of Industrial Engineering, University of Iowa

**Position title:** Professor, Department of Industrial Engineering, Ajou University

**Areas of interest:** Ergonomics, Biomedical Informatics