

# 고차원 전사체 자료를 이용한 췌장암 환자의 예후 예측 모형<sup>†</sup>

정석호<sup>1</sup> · 목리디아<sup>2</sup> · 박태성<sup>3</sup>

<sup>1,3</sup> 서울대학교 통계학과 · <sup>2,3</sup> 서울대학교 생물정보학 협동과정

접수 2018년 10월 11일, 수정 2018년 10월 12일, 게재 확정 2018년 11월 19일

## 요약

췌장암은 사망위험이 높은 대표적인 질환이며, 임상변수만으로 예후에 대한 예측이 어려워 유전적인 특성을 고려한 연구가 필요하다. 이를 위해 임상연구와 함께 유전적 연구를 바탕으로 한 예측 모형을 개발하려는 시도들이 진행되고 있다. 그런데, 최근에 관심 받고 있는 차세대 유전체 분석인 RNA 시퀀싱 발현 자료의 경우, 변수의 개수가 수만 개에 이르는 고차원 자료로써 변수의 수가 표본 수보다 훨씬 더 큰 문제가 있다. 본 연구에서는 이러한 고차원 RNA 시퀀싱 자료를 임상자료와 함께 통합하여 예후 예측을 위한 통계모형을 개발하기 위하여 (1) 유전자 필터링, (2) 후보 유전자 마커 선택, (3) 별점화 Cox 모형을 이용한 최종 마커 선택의 단계별로 모형을 개발하였다. 본 연구에서 소개할 모형 개발 방법은 RNA 시퀀싱 자료에 기반한 타 암종에 대한 예후 예측 모형의 개발을 위한 가이드라인으로 널리 활용될 것으로 기대한다.

주요 용어: 예후 예측, Cox regression, Lasso penalty, Elastic net 모형.

## 1. 서론

암과 같은 복합성 질병에 대한 환자의 유전적인 배경 연구는 지속해서 이루어져 왔다. 유전 마커를 이용한 진단이나 예측을 위한 전통적인 시도들에는 세포에서 전사체의 발현 정도를 측정한 마이크로 어레이 기법을 통해 진단이나 예후에 연관성 (association)을 가지는 특이적인 differentially expressed gene (DEG)를 찾는 연구들이 있었다 (Takikita 등, 2009; Dillhoff 등, 2008). 생물학적 실험 기법이 발전함에 따라 차세대 RNA 시퀀싱 기술을 통한 다차원 전사체 자료는 마이크로 어레이 자료에서 발생하는 외부잡음 (background noise)이 없고 유전자 탐침을 이용한 것보다 더 많은 유전자 마커를 포함하고 있다는 점에서 관심을 받고 있다. 그러나 정수자료로 이루어진 RNA 시퀀싱 자료의 특성상 기존 마이크로 어레이 자료를 이용할 때와는 다른 새로운 예측 모형이 필요하다.

암의 예후에 관한 예측 연구와 생존 분석은 꾸준히 이루어져 왔다 (Kim 등, 2016; Lee 등, 2017). 여러 암종 중에서 췌장암은 예후가 좋지 않은 대표적인 암종으로 알려져 있다. 진단을 받는 시점에서 수술적으로 제거 가능한 환자들이 20% 미만이며, 수술 후 5년 생존율 또한 25% 미만으로 현저하게 낮다 (Ryu 등, 2015). 이에 따라 환자들의 예후를 예측하기 위한 임상적인 연구가 꾸준히 이루어져 왔으나

<sup>†</sup> 이 연구는 정부(보건복지부)의 재원으로 한국보건산업진흥원(KHIDI)을 통해 보건의료기술연구개발사업 지원에 의하여 이루어진 것임 (과제고유번호 : HI16C2037).

<sup>1</sup> (08826) 서울시 관악구 관악로 1 서울대학교 통계학과, 석사과정.

<sup>2</sup> (08826) 서울시 관악구 관악로 1 서울대학교 생물정보학 협동과정, 석사과정.

<sup>3</sup> 교신저자: (08826) 서울시 관악구 관악로 1 서울대학교 통계학과, 교수.

E-mail: tspark@stats.snu.ac.kr

임상적인 요인만으로는 예후에 대해 정확한 예측을 하기는 쉽지 않다. 따라서 임상적인 요인과 더불어 췌장암의 예후에 관여하는 특이적인 유전자를 찾는 것이 중요한 문제가 되었다. 또한, 의료 중재를 통한 실제 임상에 적용하기 위해서는 단순히 예측력이 높은 모형뿐만 아니라, 예측과 관련된 인자를 잘 구분해야 할 필요가 있다.

RNA 시퀀싱 발현 자료의 경우 고차원 자료로서 많은 변수 대비 상대적으로 표본 수가 작은 문제가 있으므로 전통적인 통계 분석 방법을 적용하기 어렵다. 그러므로 본 연구에서는 고차원 자료 특성에 맞추어 예후에 중요한 역할을 하는 유전자를 찾는 과정과 이를 바탕으로 만든 유전자 예측 모형을 설계하고 모형을 평가하는 절차를 진행하였다. 이러한 절차에 기반하여 고차원 전사체 자료를 이용하여 분석하고자 할 때 고려해야 하는 요인 중에서 예측 모형에 유의미하게 영향을 줄 수 있는 인자들을 찾고, 전사체 자료를 이용한 예후 예측 성능을 높일 수 있는 분석 접근 방법을 제시하는 것이 본 연구의 목적이 다.

본 연구에서는 췌장암 환자, 특히 예후가 좋지 않은 췌장관 세포암 (Pancreatic ductal adenocarcinoma, PDAC) 환자들에 대해 RNA 시퀀싱 발현 자료와 임상 정보를 통합하여 예후를 예측하는 모형을 개발하는 구체적인 절차를 제안하였다. 이 절차를 기반으로 타 암종에 대한 예후 예측 모형의 개발을 위한 가이드라인을 제시하고자 한다.

## 2. 자료 및 분석 방법

### 2.1. 연구 자료

본 연구에서 활용한 자료는 The cancer genome Atlas (TCGA)의 genomic data commons (GDC) 포털로부터 받은 RNA 시퀀싱 자료이다. (<https://portal.gdc.cancer.gov>). NIH의 TCGA는 GDC 포털을 통해 mRNA의 발현량, 임상 정보 외에도 조직학 슬라이드 이미지 문자 정보, CpG 메틸레이션 정보, DNA copy number, miRNA 발현량 등의 데이터를 제공한다. 특별히 RNA 시퀀싱 발현량 자료의 경우, 연구 대상의 표본에서 대량의 유전자 발현 상황을 탐색하는데에 효율적이기 때문에 예후에 영향을 미치는 유전자를 탐색하기 위한 정보로써 사용하였다.

PDAC 160명의 표본은 HTseq RNA 시퀀싱 발현 자료와 환자들의 임상 정보로 이루어져 있다. 이 때, 관측 기간이 3개월 이하의 자료를 제외하고, mRNA 발현 값이 존재하는 112명에 대해서 분석을 진행하였다. RNA 시퀀싱 발현 자료의 경우 발현 정보를 fragments per kilobase million (FPKM) 방법으로 정규화시킨 값을 이용하였다 (Mortazavi, 2008). 유전자 정보의 경우 앙상블 ID 정보로 주어지기 때문에 초기 60483개의 유전자 앙상블 ID 중에서 HUGO gene nomenclature committee (HGNC)에서 지정하는 유전자로 일치하는 56480개의 유전자를 분석에 활용할 변수로 선택하였다 (Wain 등, 2002).

생존 시간의 경우 췌장암 환자가 진단받은 시점을 기준으로 사망할 때까지의 기간을 의미한다. 임상 자료는 결측값의 비율이 8% 이내인 변수를 사용하였고, TCGA에서 제공하는 변수 중에서 예후에 영향을 미치는 것으로 판단되는 변수를 일차적으로 의학적 자문을 통하여 선별하였다. Table 2.1은 임상 변수에 대한 요약된 정보가 정리되어 있다. 선별된 변수를 가지고 예후 예측 모형을 설계할 때 유전체 분석과 동시에 고려할 생존시간에 영향을 주는 임상 변수가 있는지를 확인하기 위해 전진 선택법 (forward selection), 후진 소거 (backward elimination), 단계식 변수 선택 (stepwise variable selection) 방법을 이용하여 임상 변수를 선택하였다. 단계식 변수 선택법은 초기 모형 설계에 따라 각각 전진 삽입과 후진 소거 방법의 결과와 같았다. 변수 선택 기준으로는 Akaike's information criteria (AIC)를 이용하였다 (Burnham 등, 2002). 이에 따라 최종 변수로 수술 후 잔류 종양, 수술 방법, 병변의 위치, 성별, 진단 시의 나이, 임파선 수, 종양 크기 등을 선택하였다.

생존 기간의 경우에는 PDAC 환자 총 표본 112명 중 사망한 환자 44명으로, 중도절단 비율은 61%이 있으며, 생존 기간의 중앙값은 652일이다. 병기로 살펴보았을 때, AJCC 7판 기준으로 1기에 있는 사람이 86명으로 76%를 차지하고 있고, 2기 이상의 환자는 2기와 같이 취급하여 분류하였다 (Edge 등, 2010).

**Table 2.1** Clinical variable descriptive statistics

Clinical variable	Variable description	Missing Rate	Descriptive statistics
Age	Age at diagnosis	0	64.48(11.17) mean(se)
Sex		0	Male: 64, Female: 48
Location of tumor		0	Head: 86 Body: 9 Tail: 9
Operation		0.9	Whipple: 84 Distal Pancreatectomy: 16 Other: 32
LN number	Number of Lymphocytes	0.9	17.63(8.73)
Positive LN	Number of cancer transmitted Lymphocytes	0	2.91(3.28)
Size	Maximum tumor size	0	3.914(1.66)
Residual tumor	Residual tumor after surgery	7.14	R0: 62, R1: 34 R2: 5, Other: 3
Smoking	1: No history of smoking 2: Smoking 3: History of smoking	9.8	1: 46, 2: 13, 3: 42
Alcohol diabetes		3.5 8.9	Yes: 73, No: 34 Yes: 27, No: 75
Chronic pancreatitis		9.8	Yes: 13, No: 88
Radiation therapy	Radiation therapy after surgery	8.03	Yes: 29, No: 73
Chemotherapy	Chemotherapy after surgery	8.92	Yes: 74, No: 28
Race		2.67	Asian: 9 Black or African American: 2 White: 98
Status		0	Alive: 68, Dead: 44
Overall survival time		0	Median: 652 days

## 2.2. 연구 설계

분석은 크게 (1) 유전자 필터링, (2) 후보 유전자 마커 선택 (3) 벌점화 Cox 모형의 3단계로 진행하였다. RNA 시퀀싱 자료의 경우 변수의 수가 표본의 수보다 현격히 많다. 그러므로 일반화 선형모형 등의 기존방법을 사용하기 어렵다. 또한, 변수가 6만 개 이상 존재하고 유의한 변수 역시 찾을 수 있어야 하는 점에서 앙상블, 의사결정나무와 같은 데이터마이닝 기법을 활용하기 어렵다. 따라서 이번 연구에서는 이런 기법들의 한계점을 고려하여 예측 모형을 설계하는 자세한 필터링 및 전체적인 분석과정은 Figure 2.1에 제시하였다.

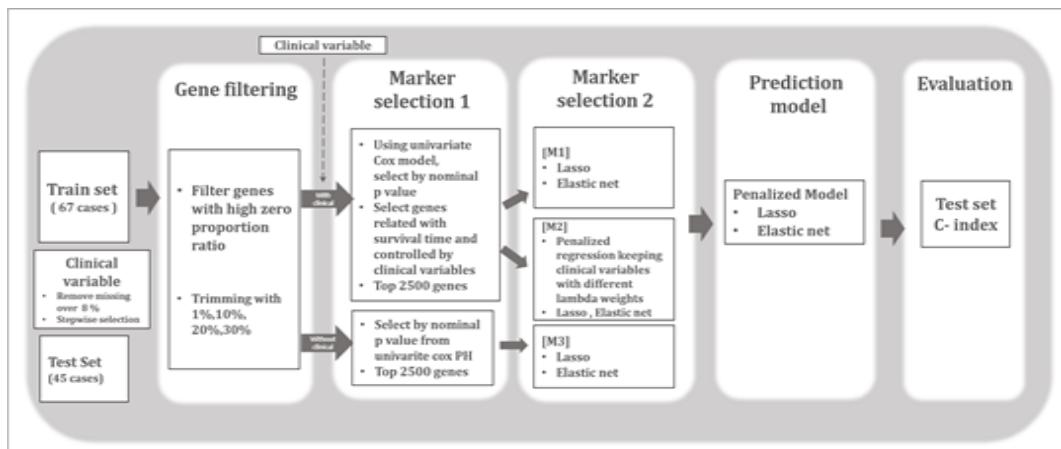


Figure 2.1 Overall scheme

### 2.2.1. 유전자 필터링

RNA 시퀀싱 자료 중 표본에 대해 전반적으로 낮은 발현 값을 보이는 유전자를 필터링 (유전자 필터링)하는 과정은 두 가지 측면에서 필요하다. 첫 번째로, 환자 개별 특성으로 나타날 수 있는 거짓 양성을 조절해 준다 (Bourgon과 Gentleman, 2010). 또한, 시퀀싱 자료 정제 과정 중 임의로 생길 수 있는 잡음을 걸러내 준다 (Sha와 Phan, 2016). 그러므로 이 논문에서는, 발현량이 거의 측정되지 않은 유전자 마커들을 필터링하는 비율을 지정하였다. 기존 연구에서는 80% 이상의 발현량이 0인 유전자 마커들을 없애는 필터링을 진행하지만, 특별히 정해진 비율은 명시하지 않았다 (Grimes 등, 2018). 그러므로 본 연구에서는 적절한 필터링 기준의 제시를 위해 발현량이 0인 유전자의 비율이 99%, 90%, 80%, 70%, 총 4가지 다른 비율로 필터링을 진행하였다.

### 2.2.2. 후보 유전자 마커 선택

필터링 이후 이를 고려한 분석의 용이성 및 차원축소를 위해 후보 마커 선택을 우선적으로 진행하였다. 임상변수를 분석에 활용한 방법에 따라 크게 다음 두 가지이다. 두 가지 분석 모두 생존 시간에 유의성을 가지는 마커를 일차적으로 선별하기 위해서 Cox의 비례위험모형을 사용하였고, 다음과 같은 식으로 나타난다. 이때, 의 확률 모형을 지정해 주지 않아도 되므로, 모형의 가정에 구애받지 않는다는 장점이 있다.

$$h(t; x) = h_0(t) \exp(\beta_1 X_1 + \cdots + \beta_p X_p). \quad (2.1)$$

본 연구에서는 후보 유전자 마커를 선택하기 위해 두 가지 방법을 고려하였다. 첫 번째는 임상 정보 변수를 공변량으로 고정하고 개별 유전자 마커들을 다변량 Cox 모형에 적합하여 유의확률을 기준으로 상위 마커들을 선택하였다. 두 번째는 임상 정보 변수를 고려하지 않고 개별 유전자 마커들을 단일 Cox 모형에 적합하여 유의확률 기준으로 상위 마커들을 선택하였다. 이 과정 중 생길 수 있는 각종 검정 문제를 해결하기 위해 유의확률 대신에 false discovery rate (FDR)를 이용하여  $q$  value를 기준으로 선택한다 (Benjamini와 Hochberg, 1995). 하지만 본 방법에서는 일차적인 마커 선택 후 Elastic Net을 이용하여 마커를 선택하는 과정을 추가로 진행하므로 후보 마커를 더 늘리기 위해 Wald 검정 통계량의 유

의학률을 기준으로 한 마커만을 선택하였다. 이때, 정확한 성능 비교 기준을 위해 유의 수준 내에서 선택되는 마커의 개수를 고정하였다. 몇 가지 유전자 필터링 기준에 따라 유의 수준 5% 이내의 마커 개수에는 차이가 있었기 때문에, 각각의 필터링 기준에 따라 통계적으로 유의한 유전자를 모두 포함하도록 유의학률을 기준으로 개수를 고정하여 일차적인 마커 선택을 진행하였다. 이후 선정된 마커를 가지고 임상 정보를 활용하는 두 가지 접근 방법을 고려하여 최종 마커 선택을 진행하였다.

### 2.2.3. 벌점화 Cox 모형

첫 번째 최종 마커 선택 방법은 위의 과정에서 찾은 후보 마커를 모두 Cox 비례위험모형을 그대로 적용할 경우 변수의 수가 표본 수 대비 매우 많아 다중 공선성 및 적합되지 않는 문제가 있으므로 변수선택과 축소 (shrinkage)를 동시에 진행하는 벌점화 기법을 이용하였다. 이때,  $L_1$ ,  $L_2$  norm을 모두 적용한 Elastic Net 방법을 생존 분석 자료에 적용하여 마커 선택을 진행하였다. 이에 따라 식 (2.2) 와 같은  $L_1$ ,  $L_2$  norm 벌점화 함수를 고려한다. 여기서  $\lambda$ ,  $\alpha$ 는 조정계수 (tuning parameter)이다.

$$P_{\lambda,\alpha}(\hat{\beta}) = \sum_{j=1}^p \lambda(\alpha|\beta_j| + 0.5(1-\alpha)\beta_j^2). \quad (2.2)$$

그 후 다음과 같이 부분 가능도 추정량 (partial Likelihood)을 통해 회귀계수들을 추정한다.

$$\hat{\beta} = \arg \min \frac{1}{n} \left[ \sum_{s=1}^S -X_{i_s}^t \beta + \log \left( \sum_{i \in R_s} \exp(X_{i_s}^t \beta) \right) \right] + P_{\lambda,\alpha}(\beta). \quad (2.3)$$

이때,  $\alpha$ 가 1인 경우는 LASSO 모형이 되며 ( $L_1$  penalty만 존재),  $\alpha$ 가 0인 경우는 Ridge 모형 ( $L_2$  penalty만 존재)가 된다. 조정계수  $\lambda$ 는 앞서 만들었던 training set에 대해 leave one out cross validation (LOOCV)를 진행하여 설정하였다. 조정계수는 validation set의 validation error를 최소로 하도록 선택하였다. 벌점화 Cox 모형을 적합할 때 후보 마커 선택과정에서 임상변수를 추가했다면 임상변수를 모형에 포함하였고 그렇지 않으면 포함하지 않았다.

두 번째 방법을 설명하기에 앞서, LASSO 및 Elastic Net 방법의 경우, 높은 상관관계를 가진 변수 중 하나만이 선택되는 문제가 존재한다 (Tibshirani, 2012). 그러므로 RNA 시퀀싱 자료와 같이 고차원 자료에서 벌점화 Cox 위험모형을 적용하는 경우 모형의 식별 가능성에 유의할 필요가 있다.

이를 해결할 수 있는 조정 (adaptive) LASSO 방법 등이 제시되어 있으나, 해석 가능한 예측 모형을 만들기 위해 본 논문에서는 앞서 선택된 변수에 더 작은 벌점 조정계수를 적용하도록 설정하였다 (Zhang 등, 2007). 모형 1에서 제시되었던 것과 달리 다음과 같이  $k$ 개의 임상변수를 고려한  $L_1$ ,  $L_2$  norm 벌점화 함수를 적용하였다.

$$P_{\lambda,\alpha}(\hat{\beta}) = \sum_{j=1}^k \lambda\gamma(\alpha|\beta_j| + 0.5(1-\alpha)\beta_j^2) + \sum_{j=k+1}^p \lambda(\alpha|\beta_j| + 0.5(1-\alpha)\beta_j^2), \quad 0 \leq \gamma < 1. \quad (2.4)$$

이를 실제로 적용한 모형에서는 임상 정보와 함께, 임상 정보 변수를 조정하여 적합한 Cox 위험모형을 통해 추정된 Wald 검정 통계량의 유의학률을 이용해 선택한 후보 마커를 이용하여 분석하였다. 이때, Elastic net과 위에서 제시된 방법의 장점을 활용하기 위해 다양한  $\alpha$ 와 임상변수에 대한 벌점계수 (penalty factor), 즉  $\gamma$ 를 이용하여 성능을 측정하였다. 이때, 가장 좋은  $\alpha$ ,  $\lambda$ 의 조합으로 (0.3, 0.3)이 선택되었다.

또한,  $\lambda$ 의 경우 앞선 벌점화 모형과 같이 LOOCV를 적용하여 변수선택과정에서 안정적인 결과를 얻을 수 있도록 하였다.

#### 2.2.4. 적용된 모형 소개

분석의 3가지 단계에서 후보 마커를 선택하는 방법, 최종 마커를 선택하는 방법을 달리하여 Figure 2.1에서 [M1], [M2], [M3]의 총 3가지 후보 모형들을 적합하였다. 모형 M1의 경우 임상변수를 고정하고 각 단일 유전자 마커들에 Cox 모형을 적합하여 얻은 Wald 검정 통계량의 유의확률을 기준으로 유의한 마커들을 후보 마커로 찾은 후 임상변수를 포함한 벌점화 Cox 모형을 적용하여 최종 마커 및 예측 모형을 찾아냈다. 모형 M2의 경우는 후보 마커를 찾는 방법은 동일하지만, 임상변수와 유전자 마커에 서로 다른 벌점을 적용하여 임상변수에 더 높은 비중을 주는 벌점화 Cox 모형을 고려하였다. 마지막으로 모형 M3의 경우, 임상변수를 따로 고정하지 않고 단변량 Cox 모형을 적합하여 유의한 후보 마커를 찾고 후보 마커를 이용, 벌점화 Cox 모형을 적용하여 최종 마커와 예측 모형을 찾았다. 즉, 모형 M1과 M2는 임상변수를 후보 마커를 찾기 이전에 고려한 점에서 M3와 차이가 있고, 모형 M2는 임상변수에 더 높은 비중을 두고 벌점화 Cox 모형을 적용한 점에서 다른 두 모형과 차이를 보여주었다.

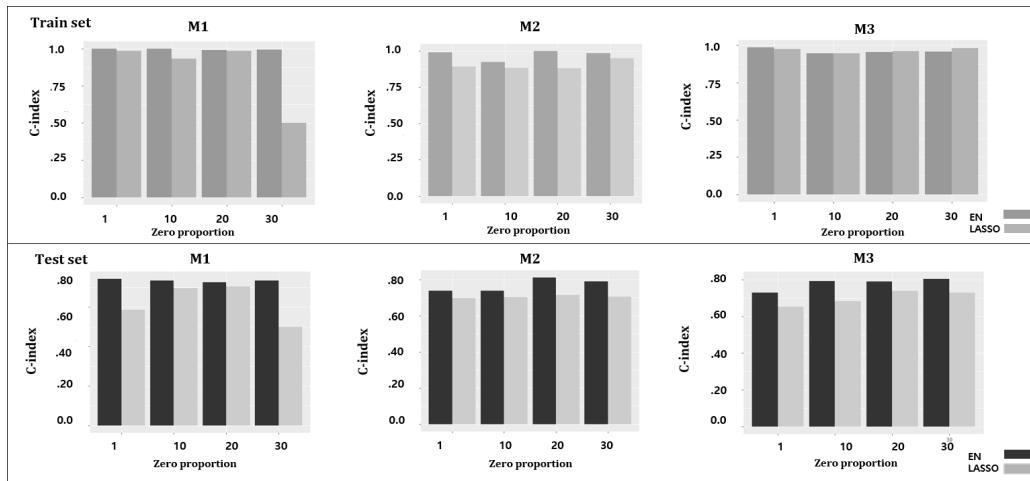
### 3. 연구 결과

전체 PDAC 환자 112명을 2:1의 비율로 나누어 모형 설계를 위한 training set과 모형 평가를 위한 test set과 분리하였다. 표본 수가 충분하지 않아 (56명) 주로 사용되는 1:1 비율 대신 2:1을 적용하였다. 이후 앞서 설명한 총 3개의 예측 모형에 적합하여 Harrell's C-index를 기준으로 모형을 평가하였다 (Harrell 등, 1996). C-index의 경우 예측 생존 시간과 실제 생존시간에 대한 일치율 (concordance)을 계산하며, 이때 예측 생존 시간에 대해 상대적인 순위 (relative rank)를 고려한다. 1에 가까울수록 높은 일치율을 의미한다.

$$\hat{C}_H = \frac{\sum_{i \neq j} \delta_i I(\hat{y}_i < \hat{y}_j) I(y_i < y_j)}{\sum_{i \neq j} \delta_i I(y_i < y_j)}, \quad 0 \leq \hat{C}_H \leq 1. \quad (3.1)$$

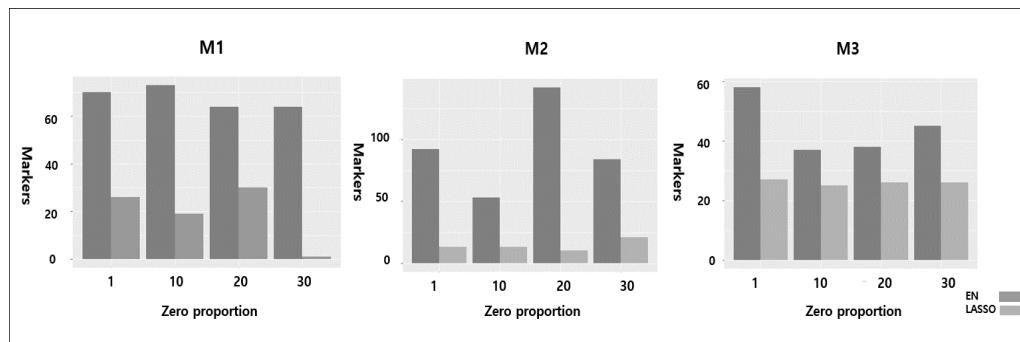
RNA 시퀀싱 자료의 경우 모든 변수에 대해 결측치가 존재하지 않으나 임상변수의 경우 결측치에 대한 고려가 필요하다. 따라서 결측치가 8% 미만이며, 단계적 삽입 방법으로 선택된 8개의 임상변수의 정보가 손실이 없는 최대 67명으로 구성된 training set으로 모형을 설계하였다. test set의 45명 중 15명의 경우 8개의 임상변수에 대한 부분적인 결측치가 존재하지만, 모형 설계와 독립적으로 볼 수 있는 test set으로 모형을 평가하는 과정을 진행하였다.

필터링 과정의 경우, 유전자 발현 정보에 대해 99%, 90%, 80%, 70%, 총 4가지 다른 비율을 설정하여 필터링을 진행하였을 때, 필터링 이후 남은 유전자 마커 수는 각각 47,756, 38,955, 35,235, 32,786개였다. 이후 후보 마커 선택 시 개별 유전자 마커들을 단변량 또는 임상변수를 고정한 다변량 Cox 모형에 적합하여 얻었을 때 필터링 기준, 임상변수 설정에 따라 서로 다른 숫자의 Wald 검정 통계량의 유의확률 5% 기준 유의 유전자 마커를 얻었다. 그러므로 정확한 성능 비교를 위해 각 설정에서 찾은 유의확률 5%에서 유의한 유전자를 후보 마커에 모두 포함할 수 있도록 2500개를 선택하였다. 그 후 벌점화 Cox 모형까지 적용하여 각 모형의 test C-index를 계산하였다. 모형별 training set과 C-index는 Figure 3.1에 나타난 바와 같다. Figure 3.1의 각 그래프에서 x축의 zero proportion은 표본 중 해당 유전자의 0이 아닌 값의 비율을 각각 1%, 10%, 20%, 30%로 설정하여 필터링한 과정을 나타내며, y축은 앞서 설명한 Harrell's C-index를 나타낸 것으로 test set의 C-index가 높을수록 좋은 성능을 보여준다.



**Figure 3.1** Prediction results for each model. Zero Proportion indicates the filtering threshold for sample proportion of nonzero counts for each gene

모형 M1의 방법으로 진행하였을 때의 결과들을 살펴보면, LASSO 방법의 경우 필터링 비율이 80%가 될 때까지는 안정적인 예측 성능을 보이지만, 필터링 기준이 70%가 될 때 알고리즘이 수렴하지 않아 training set과 test set에서 C-index가 0.5 수준까지 떨어지는 결과를 얻었다.



**Figure 3.2** Number of prediction markers

Figure 3.2에서 확인할 수 있듯이, LASSO의 경우 마커의 개수가 20개에서 30개 안팎으로 선택되는데, 이때 유전자 필터링 기준에 따라 중요한 예측 마커가 제거될 수 있음 시사한다. 한편, 상대적으로 Elastic Net 방법의 경우  $\alpha$ 값을 조정하기 때문에  $L_1$  penalty만 고려한 LASSO 방법과 달리  $L_2$  penalty를 고려하여 전체적으로 선택되는 변수의 개수가 늘어나고, 이에 따라 예측 마커의 수가 늘어나 LASSO 방법보다는 필터링 비율별 예측 성능에 안정성을 보인다.

또한, Elastic net의 경우 가능한 후보 마커들을 축소하면서 선택을 같이 진행하므로 변수선택 측면에서 LASSO 방법보다 더 안정적일 수 있다 (Zou 등, 2005; Zhao 등, 2006). 반면 방법 3에서는 필터링 기준이 99% 일 때, LASSO 방법에서 80% 이상의 비율일 때보다 예측 성능이 더 떨어지는 것을 확인할 수 있다.

3개의 방법 중 가장 안정적인 결과를 나타내는 모형 M1의 결과는 Table 3.2에서 자세하게 요약되어

있는데, 예측 성능뿐 아니라 마커 수 역시 필터링 비율에 영향을 크게 받지 않는 것으로 나타났다. 이는 모형 M2의 결과 및 모형 M3의 결과와 어느 정도 상반된 모습을 보여주었다.

**Table 3.1** Training and test C-index for Model 1

Filtering ratio	LASSO			Elastic Net		
	Training	Test	# of markers	Training	Test	# of markers
99%	0.984	0.586	26	0.998	0.743	70
90%	0.933	0.695	19	0.999	0.734	73
80%	0.984	0.704	30	0.991	0.725	64
70%	0.5	0.5	1	0.993	0.734	64

한편 예후 예측 성능이 높게 나타나는 유전자 예측 마커 중 S100A2의 경우, 암의 성장 및 전이 과정에 작용하여 예후에 영향을 미치는 유전적인 요인으로써 보고된 바 있다 (Ohuchida 등, 2007). 선택된 해당 유전 마커는 수 만개의 후보 유전자군에서 예측 성능을 가지고 있는 마커로 해석될 수 있고, S100A2와 같이 기존 연구에서 밝혀진 마커와 함께 아직 기작이 밝혀지지 않은 유전자 마커에 대해서도 예후 진단의 후보 유전자 마커로 해석할 수 있다. 또한, 이러한 유전자 후보 마커들은 예후 분자 진단 키트를 개발하기 위한 실험적 후보군으로 제시될 수 있다.

#### 4. 결론 및 고찰

이번 연구에서는 TCGA 포털에서 제공하는 PDAC 환자에 대한 임상 정보와 고차원 전사체 자료를 이용하였을 때 예후 예측 모형을 개발하는 연구를 수행하였다. 구체적으로 (1) 유전자 필터링, (2) 후보 유전자 마커 선택 (3) 별점화 Cox 모형 적합의 3단계에 걸쳐 모형을 개발하였다. 단계마다 다양한 모형들과 조합들을 고려하였다. 구체적으로 단변량 Cox 모형을 적합하여 마커 선택 시 임상변수의 고려 여부, 별점화 Cox 모형을 적용하여 유전 마커를 선택할 때 임상변수에 대한 조정계수의 설정이 예측 모형의 성능에 미치는 효과를 살펴보았다.

3가지 모형과 여러 조합을 종합한 결과, 임상변수를 고려하여 유전자 후보 마커를 선택하고 이를 모두 최종 마커 선택과정에서 적용한 모형 M1 중, Elastic Net 방법이 가장 안정적인 예측력을 보여주는 것으로 나타났다. 특히, 필터링 비율에 크게 영향받지 않고 안정적인 예측 성능을 보여주었다.

한편, 이 자료의 경우 모든 모형을 통틀어 유전자 필터링 기준은 20%가 이상적이라고 할 수 있다. 필터링 기준이 20% 미만일 경우 너무 많은 유전 변수를 고려하기 때문에 단변량 Cox 위험모형의 적합 시에 누적위험함수비율이 비정상적으로 높게 추정되거나, 마커 선택 시의 많은 변수 수로 인해 표본 수가 충분하지 않아 과적합 문제가 발생할 수 있다.

분석 시에 유전 마커 수에 비해 상대적으로 현격히 적은 표본 수로 분석을 진행하게 되기 때문에 거짓 양성 (false positive) 문제를 염두에 두어야 한다. 따라서 적당한 유전자 필터링 비율을 고려하여 모형 설계에 활용할 자료를 결정하는 것이 중요하다. 즉, 예후 예측 모형 개발을 위해 임상 정보를 포함한 별점화 Cox 모형을 고려하는 것이 합리적이며 마커 선택 전 필터링 비율 역시 20% 정도로 유지하는 것이 적합하다는 결론을 얻었다.

후속 연구에서는 bootstrapping 등으로 표본 수를 늘리는 방법을 고려해 볼 수 있다. 또한, 자료의 결측 비율이 높아 분석에 활용하지 못한 임상변수를 Mice R package 등을 이용한 imputation을 진행하여 PDAC의 예후에 영향을 미치는 임상 인자로써 추가적으로 고려해볼 수 있다 (Buuren 등, 2010). 또한, 예후 예측 모형 설계 시에 모수적 모형뿐 아니라 support vector machine과 같은 세미 모수적 모형과 random forest를 생존 분석에 활용한 비모수적 모형으로 예측 성능을 더 높이는 것을 기대할 수 있다 (Van Belle 등, 2010; Ishwaran 등, 2008).

본 연구에서는 PDAC 환자들에 대해 RNA 시퀀싱 발현 자료와 임상 정보를 통합하여 예후를 예측하는 모형을 개발하는 구체적인 절차를 제안하였다. 본 연구에서 시도한 방법은 암 종이나 기타 전사체 자료 특성에 크게 구애받지 않는 일반적인 접근법이었다. 따라서 이에 근거한 구체적인 예측 모형 개발 가이드라인은 다차원 전사체 자료를 이용한 다른 암종 예후 및 유전체 자료 등과의 통합 분석에 쉽게 적용될 수 있을 것이다.

## References

- Abdi, H. (2010). Holm's sequential Bonferroni procedure. *Encyclopedia of Research Design*, **1**, 1-8.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.
- Bourgon, R., Gentleman, R. and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, **107**, 9546-9551.
- Buuren, S. V. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1-68.
- Dillhoff, M., Liu, J., Frankel, W., Croce, C. and Bloomston, M. (2008). MicroRNA-21 is overexpressed in pancreatic cancer and a potential predictor of survival. *Journal of Gastrointestinal Surgery*, **12**, 2171.
- Edge, S. B. and Compton, C. C. (2010). The american joint committee on cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of Surgical Oncology*, **17**, 1471-1474.
- Fischer, L. K., Katz, M. H., Lee, S. M., Liu, L., Wang, H., Varadhachary, G. R., Fleming, J. B. et al. (2016). The number and ratio of positive lymph nodes affect pancreatic cancer patient survival after neoadjuvant therapy and pancreaticoduodenectomy. *Histopathology*, **68**, 210-220.
- Graham, E. M., Baird, A. H. and Connolly, S. R. (2008). Survival dynamics of scleractinian coral larvae and implications for dispersal. *Coral Reefs*, **27**, 529-539.
- Grimes, T., Walker, A. R., Datta, S. and Datta, S. (2018). Predicting survival times for neuroblastoma patients using RNA-seq expression profiles. *Biology Direct*, **13**, 11.
- Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361-387.
- Kim, B., Ha, I. D. and Lee, D. (2016). Analysis of multi-center bladder cancer survival data using variable-selection method of multi-level frailty models. *Journal of the Korean Data & Information Science Society*, **27**, 499-510.
- Kim, Y. and Kong, L. (2015). Estimation of C-index for cox proportional hazards model with censored biomarker covariate subject to limits of detection. *Journal of Biopharmaceutical Statistics*, **25**, 459-473.
- Lee, T. and Lee, M. (2017). Analysis of stage III proximal colon cancer using the Cox proportional hazards model. *Journal of the Korean Data & Information Science Society*, **28**, 349-359.
- Ohuchida, K., Mizumoto, K., Miyasaka, Y., Yu, J., Cui, L., Yamaguchi, H., Yamaguchi, K. et al. (2007). Over-expression of S100A2 in pancreatic cancer correlates with progression and poor prognosis. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, **213**, 275-282.
- Ryu, J. K. (2015). The early detection of pancreatic cancer: Whom and how? *Korean Journal of Pancreas and Biliary Tract*, **20**, 198-203.
- Sha, Y., Phan, J. H. and Wang, M. D. (2015). Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (pp. 6461-6464). IEEE.
- Takikita, M., Altekruse, S., Lynch, C. F., Goodman, M. T., Hernandez, B. Y., Green, M., Zeruto, C. et al. (2009). Associations between selected biomarkers and prognosis in a population-based pancreatic cancer tissue microarray. *Cancer Research*, **69**, 2950-2955.
- Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.
- Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, **7**, 1456-1490.
- Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W. and Povey, S. (2002). Guidelines for human gene nomenclature. *Genomics*, **79**, 464-470.

- Wang, W., Chen, S., Brune, K. A., Hruban, R. H., Parmigiani, G. and Klein, A. P. (2007). PancPRO: Risk assessment for individuals with a family history of pancreatic cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, **25**, 1417.
- Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 856-863.
- Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541-2563.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301-320.

## Development of prediction models using high dimensional RNA sequencing data for the prognosis of pancreatic ductal adenocarcinoma<sup>†</sup>

Seokho Jeong<sup>1</sup> · Lydia Mok<sup>2</sup> · Taesung Park<sup>3</sup>

<sup>13</sup>Department of Statistics, Seoul National University

<sup>23</sup>Interdisciplinary Program in Bioinformatics, Seoul National University

Received 11 October 2018, revised 12 October 2018, accepted 19 November 2018

### Abstract

Pancreatic cancer is a well known disease with a high risk of death. Accurate prediction of prognosis using only clinical information has not been easy. Therefore, an effort to develop a better prediction model by using genetic information along with clinical information is needed. RNA sequencing data consist of tens of thousands of gene expression variables. As a result, the number of variables is much larger than sample size. In this study, we developed the prognosis prediction model by integrating the high dimensional RNA sequencing data with clinical data through the following three steps: (1) gene filtering, (2) selecting candidate genetic markers, (3) final marker selection using penalized Cox model. The prognosis prediction model development procedure introduced in this study is expected to be widely used for the development of prognosis prediction models for other types of cancer as well.

*Keywords:* Cox regression, elastic net, prognosis prediction, lasso penalty.

<sup>†</sup> This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI16C2037).

<sup>1</sup> Graduate Student, Department of Statistics, Seoul National University, Seoul 08826, Korea.

<sup>2</sup> Graduate Student, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea.

<sup>3</sup> Corresponding author: Professor, Department of Statistics, Seoul National University Seoul 08826, Korea. E-mail: tspark@snu.ac.kr