한국보건정보통계학회지 제34권 제2호
Journal of the Korean Society of Health
Information and Health Statistics
Volume 34, Number 2, 2009, pp. 139～152

구축된 간암 발생 예측 모형의 개선방안에 관한 연구

이혜선[1], 명성민[2], 김도영[3], 한광협[3], 송기준[1]†
[1]연세대학교 의과대학 의학통계학과
[2]중원대학교 의료공학부
[3]연세대학교 의과대학 내과학교실

A study on the updating of prediction model for the development of hepatoma

Hye-Sun Lee[1], Sung-Min Myung[2], Do-Young Kim[3], Kwang-Hyub Han[3], Ki-jun Song[1]†
[1]Department of Biostatistics, Yonsei University College of Medicine
[2]Department of Medical Informatics, Jungwon University
[3]Department of Internal medicine, Yonsei University College of Medicine

Abstract

Objectives: The statistical prediction models are useful to establishing diagnostic and treatment rule in clinical area. So, there is an increasing interest in building a precise model to predict the probability of diseases for individual patient. In doing that, it is important to reflect the patient's changeable characteristics for improvement of predictive power. In this paper, we studied the methods for the updating of prediction model that add the information of new patients to the existing model.
Methods: To update the prediction model, we used an established model including 7 risk factors such as diagnostic type, hepatitic virus type, age, sex, α-FP, ALT, and drinking history and did the re-calibration and shrinkage of intercept and slope of existing one.
Results: we considered 4 updating methods, that is, the first one is to use existing model as it is and the second one is to re-calibrate the overall intercept. Also the third one is to re-calibrate overall intercept and slope and the last one is to re-calibrate and shrink overall intercept, and individual slope.
Conclusions: Updating methods contain old and new informations. And the model updating method by using many data can be improved predictive power. Especially, the last updating method was found to be the most accurate and useful one.

Key Word: prediction model, update method, re-calibration, shrinkage, intercept, slope

* 본 논문은 보건복지가족부 보건의료기술연구개발 사업(A050021)의 지원에 의해 이루어진 것임.
† 교신저자 : 송기준, 서울특별시 서대문구 신촌동 134 연세대학교 의과대학 의학통계학과
E-mail : biostat@yuhs.ac

1. 서 론

1.1 연구 배경 및 목적

사망률이 가장 많이 증가한 사인은 암이며, 암 사망률은 인구 10만명당 사망자수로 점차 증가하고 있다. 그리고 인구의 급격한 고령화에 따라 암 환자 발생 및 사망이 더욱 증가될 것으로 예상된다(통계청, 2008). [Table 1]은 암 사망률 현황으로 2008년 암에 의한 사망률(인구 10만명당)은 폐암(29.9%), 간암(22.9%), 위암(20.9%), 대장암(13.9%) 순으로 높은 것을 알 수 있다.

Table 1. Cancer mortality

		2003	2004	2005	2006	2007	2008
Stomach cancer	no. of deaths	11,701	11,190	10,935	10,716	10,563	10,312
	mortality rates	24.2	23.1	22.5	21.9	21.5	20.9
Lung cancer	no. of deaths	12,673	13,246	13,733	14,027	14,278	14,791
	mortality rates	26.2	27.3	28.2	28.7	29.1	29.9
Liver cancer	no. of deaths	10,916	10,861	10,877	10,884	11,144	11,292
	mortality rates	22.6	22.4	22.3	22.3	22.7	22.9
Colon cancer	no. of deaths	5,484	5,859	6,043	6,244	6,650	6,855
	mortality rates	11.4	12.1	12.4	12.8	13.5	13.9
Breast cancer	no. of deaths	1,404	1,484	1,573	1,598	1,670	1,731
	mortality rates	2.9	3.1	3.3	3.3	3.4	3.5
Uterine cancer	no. of deaths	1,397	1,325	1,345	1,240	1,241	1,261
	motarlity rates	2.9	2.7	2.8	2.5	2.5	2.5
All others	no. of deaths	19,757	20,334	20,595	20,793	22,007	22,670
	mortality rates	40.9	41.9	42.3	42.5	44.8	45.9

이 중 두 번째로 사망률이 높은 간암은 한국에서 흔한 암 중 하나로 진행된 상태에서 발견된 경우는 효과적인 치료가 어려우며 대부분이 6개월 이내에 사망하기 때문에 현재로서는 조기 진단이 간암의 생존률을 높일 수 있는 효과적인 방법이다[1]. 정기적인 초음파검사 및 α-fetoprotein(이하 α-FP) 등의 선별 검사를 이용한 조기 진단의 중요성이 강조되고 있다. 간암 고위험군에서 간암의 조기 진단을 위한 선별 검사는 간암의 효과적인 치료와 생존기간의 향상을 위하여 매우 중요하므로 간암의 조기 진단을 위한 예측모형의 구축은 중요한 이슈로 나타나고 있다.

의학 분야에서 예측모형은 환자의 위험 요인을 토대로 환자의 질병을 예측함으로써 진단과 치료의 의사결정에 도움을 줄 수 있다[2][3]. 예측모형은 예측모형 구축 시 사용된 환자 자료 외에 유사한 형태의 다른 자료를 예측하는 것 역시 가능하다. 예측 모형은 미래 환자들에 대한 예측력을 향상시키기 위해 변화하는 환자의 양상을 반영하여 개선시킬 필요성이 있다.

새로운 모형을 다시 만들지 않고, 기존 환자 정보를 반영한 예측모형에 새로운 환자 정보를 추가함으로써 모형을 개선시켜 예측력을 향상시키고자 한다. 본 연구에서는 모형을 개선시키는 방법으로 Steyerberg(2004)에 의해서 제안된 4가지 방법을 비교 평가하고 이 중 가장 좋은 방법을 제시하여 향후 미래 환자를 보다 정확히 예측함으로써 환자의 진단과 치료의 의사결정에 도움을 주고자 한다.

1.2 연구 내용 및 방법

본 연구는 최근 환자 자료 분석 시 기존 자료의 연구 결과 간암 발생 위험인자라고 알려진 간경변, 만성 간염, C형 간염, B형 간염, 연령, 성별, α-FP, ALT, 상습적 음주자, 미확인 음주력을 포함한 10개의 변수들을 적용하여 기존에 만들어진 예측모형의 절편과 기울기를 재보정과 축소 방법

을 이용하여 기존 예측모형을 개선시킨다. Steyerberg(2004)에 의해서 4가지 방법이 제안되었는데, 첫째로 기존 식을 개선하지 않고 기존 식의 절편과 기울기 사용하는 방법이 있다. 둘째로 절편만을 재보정, 셋째로 절편과 전체적인 기울기를 재보정 하는 방법이 있다. 넷째로 절충 모형으로 절편과 전체적인 기울기와 개별 위험인자의 기울기를 재보정하고 축소시키는 방법이 있다. 본 연구에서는 위의 4가지 방법으로 예측모형을 개선시키고 예측력을 비교 평가하여 최종적으로 가장 예측력이 뛰어난 모형을 구축하고자 한다.

2. 이론적 배경

2.1 로지스틱 회귀분석

로지스틱 회귀분석은 선형 회귀분석과는 다르게 종속변수가 두 개의 범주로 이루어진 경우에 독립변수와의 관계를 살펴보기 위해 사용된다. 일반적으로 종속변수가 취할 수 있는 값은 어떤 사건이 발생된 경우를 1로 하고, 발생되지 않는 경우를 0으로 하여 독립변수가 주어졌을 때 사건발생의 조건부 확률을 로짓 변환하여 사용한다.
종속변수는 위와 같은 이분형 변수이고, 독립변수는 연속형 변수와 범주형 변수를 합쳐 p개가 있다고 가정하자. 이 때 p개의 독립변수에 대한 종속변수가 1을 가질 확률을 $P(Y=1|x_1,x_2,\cdots,x_p)$라고 하고, 로짓 변환 하면,

$$\ln\left[\frac{P(Y=1|x_1,x_2,\cdots,x_p)}{1-P(Y=1|x_1,x_2,\cdots,x_p)}\right]=$$

$$\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$$

이다. 이를 로지스틱 회귀모형이라 하며, 특히 독립변수가 둘 이상인 경우를 다중 로지스틱 회귀모형이라 한다. 이 식을 $P(Y=1|x_1,x_2,\cdots,x_p)$에 관하여 정리하면,

$$P(Y=1|x_1,x_2,\cdots,x_p)=$$

$$\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}$$

이고, 로지스틱 반응함수라고 한다.

로지스틱 회귀모형에서 회귀계수를 추정하기 위해 최대우도법을 이용한다. 이 때 우도는 관찰된 자료가 발생될 확률을 알려지지 않은 모수들의 함수로 표현한 것이며, 이 우도를 최대로 하는 회귀계수를 추정하는 방법이 최대우도법이다. 이렇게 얻어진 모수들의 추정량을 최대우도추정량이라 한다.

우도함수의 최대값은 미분을 통해 얻게 되는데, 회귀계수의 최대 우도 추정값은 비선형이므로 피셔의 스코어링방법이나 뉴튼-랩슨 방법 등과 같은 반복적인 추정방법에 의하여 근사값을 구한다. 반복적인 추정 절차는 특정한 수렴기준을 만족할 때까지 계속하게 된다.
이 중 모수 β를 반복적으로 추정하는 절차를 이용한 뉴튼-랩슨 방법에 대해 간단히 소개하자면 식은 다음과 같다.
로그우도함수,

$$\ln L(\beta)=\sum_{i=1}^{n}[y_i\ \ln(p)+(1-y_i)\ln(1-p)]$$

를 구한다. 여기서

$$p=\frac{1}{1+\exp\{-(\beta_0+\beta_1X)\}}$$

이다. 이를 β_0와 β_1에 대해 편미분하여 정규방정식을 구하면,

$$\sum_{i=1}^{n}(y_i-p)=0\quad,\quad\sum_{i=1}^{n}x(y_i-p)=0$$

이고, 뉴튼-랩슨 방법을 이용하여 최대우도추정

량을 구할 수 있다. 뉴튼-랩슨은 식은

$$\beta^{(t+1)} = \beta^{(t)} - \left(\frac{\partial^2 [\ln L(\beta)]}{\partial \beta^{(t)} \partial \beta^{'(t)}} \right)^{-1} \left(\frac{\partial [\ln L(\beta)]}{\partial \beta^{(t)}} \right)$$

이다. 반복절차에서 $(t+1)$-번째에서 구한 모수벡터 $\beta^{(t+1)}$과 t-번째에서 구한 모수벡터 $\beta^{(t)}$ 사이의 차이가 설정한 기준보다 작다면 반복절차를 중단하고 마지막 단계에서 구한 값이 최대우도 추정값 $\hat{\beta}$이 된다.

회귀계수를 추정한 후에 로지스틱 회귀모형에 대한 검정은 우도비 검정, 왈드 검정, 스코어 검정 등을 이용한다.

독립변수가 많은 경우에는 변수선택법을 이용하여 로지스틱 회귀모형의 독립변수를 선택 할 수 있다. 변수를 선택하는 방법에는 설명변수의 각각의 기여도에 따라 단계별로 하나씩 추가하면서 변수를 선택하는 전진선택법과 모든 독립변수를 포함한 완전모형에서 불필요한 변수를 단계별로 하나씩 제거해 나가는 후진 제거법, 각 단계에서 변수의 선택과 제거를 반복하면서 독립변수를 결정하는 단계적 선택법이 있다.

구축된 로지스틱 회귀모형의 적합도를 보기 위한 방법으로 잔차에 기초한 χ^2 통계량과 호즈머-렘쇼(Hosmer-Lemeshow)의 합도 검정이 있다. 이 중 잔차에 기초한 χ^2 통계량은 피어슨 χ^2 통계량, 데비언스 χ^2 통계량이 있다[4,5,6,7].

2.2 로지스틱 회귀모형의 재보정(re-calibration)

기존 자료로부터 Y에 대한 예측 모형은

$$\tilde{Y}_{pred} = \tilde{\alpha}_{model} + \tilde{\beta}_1 X_1 + \cdots + \tilde{\beta}_p X_p$$

이다. 이 모형을 하나의 선형 예측인자로 가정하며 Z라 하고,

$$Z = \tilde{\alpha}_{model} + \tilde{\beta}_1 X_1 + \cdots + \tilde{\beta}_p X_p$$

로 정의된다. 위의 Z를 사용하여 회귀모형에 적합 시켜 새로운 회귀계수를 추정하는 것을 재보정(re-calibration)이라 한다. 재보정식은 새로운 회귀계수 $\hat{\alpha}_{overall}$, $\hat{\beta}_{overall}$을 추정하는 식으로 $\hat{Y}_{cal}$은

$$\hat{Y}_{cal} = \hat{\alpha}_{overall} + \hat{\beta}_{overall} Z$$

이다. 여기서 $\hat{\alpha}_{overall} = 0$, $\hat{\beta}_{overall} = 1$인 경우 기존 자료로부터 구축된 예측모형이 새로운 자료를 잘 예측하는 것으로 타당성이 높은 모형을 의미한다.

추정된 모수들은 왈드 검정통계량

$$W = \left(\frac{\hat{\beta}}{se(\hat{\beta})} \right)^2 \sim \chi_1^2, \quad where \quad W > \chi_{1,1-\alpha}^2$$

를 통해 검정될 수 있다. 만약 $\alpha_{overall}$, $\beta_{overall}$이 유의하게 얻어진다면, 이는 재보정이 필요한 모형을 의미하며, 이는

$$\hat{\beta}_{cal} = \hat{\beta}_{overall} \tilde{\beta}_i$$

로 추정된다. 재보정된 모형의 회귀계수는 기존 모형으로 얻어진 회귀계수의 상대적인 효과를 의미한다.

재보정된 회귀모형의 타당도는 F-검정을 이용하여 재보정 모형과 새로운 자료만을 적용한 모형을 비교할 수 있다. 새로운 자료만을 적용한 회귀모형은

$$\hat{Y}_{new} = \hat{\alpha}_{new} + \hat{\beta}_{1,(new)} X_1 + \cdots + \hat{\beta}_{p,(new)} X_p$$

이며, F-통계량은

$$F_{cal} = \frac{\sum(\hat{Y}_{new} - \hat{Y}_{cal})^2/(p-1)}{\sum(Y - \hat{Y}_{new})^2/(m-p-1)} \sim$$

$$F_{p-1,\, m-p-1,\, 1-\alpha}$$

이다. p는 독립변수의 수이며 m은 자료의 전체의 수를 의미하며, F-검정결과 유의하다면 새로운 자료만을 적용한 모형이 더 낫다고 통계적 의사결정을 할 수 있으며, 유의하지 않다면 재보정 모형이 더 낫다고 할 수 있다.

2.3 로지스틱 회귀모형의 축소(shrinkage)

축소(shrinkage)란 자료에서 공변량의 수가 적어도 3이상으로 존재할 때, 미래 관측치의 평균제곱 오차(Mean Square Error: MSE)를 감소시키는데 유용한 기법이다[8,9,10].
전체 평균을 줄이는 일반적인 축소모형은

$$Y_{pred} = \overline{Y} + \hat{c}\hat{\beta}_1(X_1 - \overline{X_1}) + \cdots + \hat{c}\hat{\beta}_p(X_p - \overline{X_p})$$

이다.

축소의 양을 추정하기 위해서 두 가지 방법을 고려할 수 있다. 첫째로, 붓스트랩이나 교차타당성을 이용한다[11,12]. 둘째로, 발견적 방법으로 축소요인(Heuristic shrinkage Factor: c)을 이용하며,

$$\hat{c} = (F_{model} - 1)^+ / F_{model}$$

이다. 여기서 F 통계량은 최소제곱추정법에 의해 계산된다. c의 구간을 [0,1]로 제안하고, c를 0에서 절단시킨다.

2.4 로지스틱 회귀모형의 절충 모형

위 2.2~2.3절에서 제안한 재보정과 축소모형을 이용해 절충 모형을 고려할 수 있다. 절충 모형은 기존 자료를 이용해서 새로운 자료에 적용시켜 만든 재보정 모형 $\hat{Y}_{cal}$과, 새로운 자료만을 이용해서 만든 모형 $\hat{Y}_{new}$의 가중 모형으로 식은,

$$Y_{pred} = (1 - \hat{c}_{cal})\hat{Y}_{cal} + \hat{c}_{cal}(\hat{\alpha}_{model} + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p)$$

$$= (1 - \hat{c}_{cal})\hat{Y}_{cal} + \hat{c}_{cal}\hat{Y}_{new}$$

이다. 여기서 $\hat{c}_{cal}$은 축소요인으로

$$\hat{c}_{cal} = (F_{cal} - 1)^+ / F_{cal}$$

이다.

축소는 E(MSE)를 줄이는 것으로 기대될 수 있으며, 일반화된 축소(generalized shrinkage)는 몇 개의 중요한 공변량과 나머지의 중요하지 않은 공변량으로 나눌 때 유용하다.
위의 식으로부터 재보정과 축소모형의 조합인 절충 모형의 회귀계수는 아래와 같이 추정할 수 있다.

$$\hat{\beta}_{shrunk+cal} = \hat{\beta}_{cal} + \hat{c}_{cal}(\hat{\beta}_{new} - \hat{\beta}_{cal})$$

여기서, 절충모형은 기존자료와 새로운 자료를 동시에 반영하며 공변량의 유용성을 다시 한 번 평가할 수 있으므로, 개선방안에 유용하게 적용될 수 있다.

3. 연구방법

3.1 예측 모형 개선방안

본 연구에서는 이전에 제안되었던 예측모형을 개선시키기 위하여 [Table 2]의 4가지 개선방안을 고려한다. 이 방법들은 기존 모형을 개선시키기 위해 추정된 예측인자의 수를 유지하였다.

Table 2. Updating methods

No.	Updating method	Predictors considered	Parameters estimated
1	No adjustment	10	0
2	Intercept α	10	1
3	α+ calibration slope $\beta_{overall}$	10	2
4	$\alpha+\beta_{overall}+\gamma_{1,\cdots,8\|p\le 0.05}$	10	2～12

(1) 개선방안 1

개선방안 1은 기존 예측 모형에 고정된 절편을 포함한 모든 회귀계수를 유지하는 것으로

$$Z_1 = \alpha_{old} + \sum_{i\in 1,\cdots,8} \beta_{i,old}\, x_i (\text{기존 식 유지})$$

이다. 여기서, α_{old}는 절편이고, $\beta_{i,old}$는 이전 연구에서 개발된 회귀계수이고, x_i는 새로운 연구에서 표본으로 얻어진 독립변수이다. 이 방법을 통해 얻은 식은 기존 식의 타당성을 F-검정으로 평가하며, 다른 개선방안의 참고 식으로 활용된다.

(2) 개선방안 2

재보정 방법을 이용하여 절편을 개선한 개선방안 2는

$$Z_2 = \hat{\alpha} + Z_1$$

이다. 여기서 α는 모수이고, 절편 α와 상쇄변수(offset variable) Z_1을 기반으로 새로운 자료를 이용하여 로지스틱 회귀모형에 적합 시킨다.

절편의 개선은 많은 개체의 계산을 올바르게 하며, 평균 예측 확률(average predicted probability)을 관찰된 전반적인 사건의 비율과 같게 만드는 경향이 있다.

(3) 개선방안 3

개선방안 3은 개선방안 2와 같이 재보정하는 방법이나, 개선방안 2는 절편만을 재보정하고 개선방안 3은 절편과 회귀계수를 동시에 재보정한다. 개선방안 3은

$$Z_3 = \hat{\alpha} + \hat{\beta}_{overall} Z_1$$

이다. 여기서 절편 α와 기울기 $\beta_{overall}$을 구하고, 공변량인 Z_1을 기반으로 새로운 자료를 이용하여 로지스틱 회귀모형에 적합 시킨다.

(4) 개선방안 4

개선방안 4는

$$Z_4 = \hat{\alpha} + \hat{\beta}_{overall} Z_1 + \sum_{i\in s} \hat{\gamma}_i x_i$$

$$(\hat{\gamma}_i = \hat{\beta}_i - \hat{\beta}_{overall}\beta_i)$$

이다. 여기서 s는 선택된 공변량을 가리키며, 공변량 $1,\cdots,10$ 중 최대 10개까지 선택할 수 있다. 그리고 γ_i는 재보정된 계수의 값으로부터 도출된 값으로

$$\hat{\gamma}_i = \hat{\beta}_i - \hat{\beta}_{overall}\beta_{i,old}$$

이다.

개선방안 4는 구조적으로 모형을 변화시키는 것으로 모형 수정이라고 한다.

3.2 모형 비교

개선방안들을 통해 얻은 모형의 비교 평가를 위해 보정그림과, ROC 곡선, 5가지 평가 척도를

이용한다.

첫째, 보정그림은 가로축이 예측확률, 세로축이 실제확률로 예측확률과 실제확률이 일치하는 그림일수록 예측력이 좋은 모형이라 할 수 있다.

둘째, ROC 곡선은 민감도와 특이도의 개념을 이용한 것으로 민감도는 실제 $y=1$인 자료 중에서 로지스틱 모형에 의해 옳게 예측된 비율, 특이도는 실제 $y=0$인 자료 중에서 로지스틱 모형에 의해 옳게 예측된 비율을 말한다. 민감도와 특이도가 동시에 높아야 로지스틱 모형의 예측 정도가 높다고 할 수 있다. ROC 곡선은 1-특이도에 대한 민감도의 그래프로 곡선 아래의 면적에 의해 예측정도가 평가된다. 아래 면적은 0.5와 1 사이의 값을 갖는데 면적 크기가 0.5 정도면 로지스틱 모형이 종속변수를 분별하는데 적절하지 못함을 의미하고, 0.5보다 클수록 예측 정도가 증가한다고 판단할 수 있다.

셋째, 5가지 평가 척도인 추정된 모수의 수, 보정평가 지수 U 통계량(miscalibration), c 통계량, 브라이어 점수(Brier score), 상관계수 r을 이용한다. 추정된 모수의 수는 적을수록 모형 구축 시 접근이 용이함을 나타낸다. 보정 평가 지수 U 통계량은 보정이 잘 이루어졌는지는 나타내는 척도로 잘 보정된 모형들은 0에 가까운 α와 1에 가까운 β를 갖는다. 이 통계량은 $\alpha=0$, $\beta=1$인 모형과 개선방안에 의해 개선된 식의 α, β에 해당하는 모형의 -2log likelihood의 차이를 n수로 나누어 비교한다. c 통계량은 ROC 곡선의 면적(AUC)으로 1에 가까울수록 예측력이 높은 모형을 나타낸다. 브라이어 점수(Brier score)는 전반적인 수행을 평가하는 척도로 평균 예측 오차에 관한 개념으로 $\sum(y_i-p_i)^2/n$에서 y_i는 관측된 결과, p는 각 해당 개체의 예측치를 나타낸다. 브라이어 점수는 완벽한 모형에서 0이 나타나는 척도로 이 값을 구하여 비교한다. 상관계수 r은 예측확률과 실제확률의 상관관계를 나타내는 값으로 1에 가까울수록 좋은 모형이라 할 수 있다.

4. 결 과

4.1 자료에 대한 개요

자료에 대한 개요는 [Table 3]에 제시하였다. 1990년부터 2000년까지 연세의료원에 내원하여 간암 발생 위험군으로 판단되어 2회 이상 정기적으로 복부 초음파검사를 포함한 검진을 받아온 4,339명에 대해 임상 자료를 포함한 상세 자료가 있다. 그 중 간암 발생여부를 확인할 수 있는 환자 1,827명을 연구 대상으로 하였다. 이 때 1990년 1월부터 1998년 12월까지 기존 환자 994명의 자료는 기존 예측모형 구축 시 이용되었고, 1999년 1월부터 2000년 12월까지 최근 환자 833명의 자료는 기존 예측모형을 개선시키는데 이용하였다.

이 때 기존 예측모형 구축 시 이용된 994명의 환자 자료에서 9.05%인 90명이 간암 진단을 받았고, 예측모형 개선에 이용된 최근 환자 833명의 환자 자료에서는 5.28%인 44명이 간암 진단을 받았다[13,14].

Table 3. Summary of data

	Old data	New data
Entering period	1990. 1~1998. 12	1999. 1~2000. 12
No. of total patients	994	883
No. of lung cancer patients	90(9.05%)	44(5.28%)

(1) 분석에 쓰인 자료 설명

분석을 위해 기존 예측모형 구축 시 선별된 변수인 간경변, 만성 간염, C형 간염, B형 간염, 연령, 성별, α-FP, ALT, 상습적 음주자, 미확인 음주력을 포함한 10개의 변수들을 사용하고, 이러한 변수에 대해 [Table 4]에서 정리하였다. 연속형 변수들은 이분형 변수로 바꾸어 분석에 사용하였다.

Table 4. Description of risk factor

Variable	Description	Positive
LC	Liver cirrhosis	-
CH	Chronic hepatitis	-
HCV	Hepatitis C virus	-
HBV	Hepatitis B virus	-
AGE	-	≥40
SEX	-	Male
α-FP	Alpha-fetoprotein	≥20(IU/㎖)
ALT	Alanine aminotransferase	≥40(IU/ℓ)
Heavy alcohol	-	5년 이상동안 매일 알코올 80g 이상을 섭취한 자
Unknown alcohol	-	-

분석에 사용한 기존 자료와 새로운 자료에 대한 간암 발생여부에 관한 위험요인들의 빈도표를 [Table 5]에서 제시하였다.

Table 5. Frequency table of risk factor

		Old data(n=994)	New data(n=833)
Ultrasonography	LC	335(33.7%)	282(33.85%)
	CH	540(54.33%)	460(55.22%)
	Carrier, Other	119(11.97%)	91(10.92%)
Hepatitis	HCV	121(12.17%)	133(15.97%)
	HBV	781(78.57%)	613(73.59%)
	NonBNonC	92(9.26%)	87(10.44%)
AGE	≥40	798(80.28%)	635(76.23%)
	<40	196(19.72%)	198(23.77%)
SEX	Male	683(68.71%)	568(68.19%)
	Female	311(31.29%)	265(31.81%)
α-FP	≥20(IU/㎖)	191(19.22%)	120(14.41%)
	<20(IU/㎖)	803(80.78%)	713(85.59%)
ALT	≥40(IU/ℓ)	552(55.53%)	521(62.55%)
	<40(IU/ℓ)	442(44.47%)	312(37.45%)
Drinking	Heavy alcohol	149(14.99%)	110(13.21%)
	Non/Social alcohol	543(54.63%)	628(75.39%)
	Unknown alcohol	302(30.38%)	95(11.4%)

4.2 개선방안에 따른 예측 모형

(1) 4가지 개선방안의 적용

1990년 1월부터 1998년 12월까지 기존 환자 994명의 자료를 통해 얻은 기존 예측 모형 식은 다음과 같다.

기존 예측 모형 식의 타당성을 확인하기 위해 1999년 1월부터 2000년까지 최근 환자 833명의 자료와 비교하고, 개선방안을 적용하였다.

$$\text{Risk Index(RI) for HCC}=e^{A}$$

$$\begin{aligned}A=&-6.254+(1.722\times\text{간경변})+(0.734\times\text{만성 간염})+(1.263\times\text{C형 간염})\\&+(0.775\times\text{B형 간염})+(1.315\times\text{연령}(\geq 40\text{세}))+(0.3\times\text{남성})\\&+(0.826\times\alpha\text{-FP}(\geq 20\text{IU/㎖}))+(0.283\times\text{ALT}(\geq 40\text{IU/}\ell))\\&+(0.584\times\text{상습적 음주자})+(0.222\times\text{미확인 음주력})\end{aligned}$$

$$\text{Probability for HCC}=e^{A}/(1+e^{A})$$

Table 6. Logistic regression coefficient(standard deviance) of old data and new data

Variable	Old data(n=994)	New data(n=833)
Intercept	-6.254(1.053)	-7.697(1.383)
LC	1.722(0.409)	2.324(1.038)
CH	0.734(0.265)	0.454(1.082)
HCV	1.263(0.499)	1.005(0.799)
HBV	0.775(0.396)	1.433(0.621)
AGE(≥40)	1.315(0.003)	0.816(0.564)
SEX(male)	0.300(0.328)	1.617(0.574)
α-FP(≥20IU/㎖)	0.826(0.464)	0.928(0.396)
ALT(≥40IU/ℓ)	0.283(0.150)	-0.800(0.359)
Heavy alcohol	0.584(0.393)	0.175(0.432)
Unknown alcohol	0.222(0.400)	1.327(0.448)

기존 자료와 새로운 자료를 토대로 로지스틱 회귀분석을 수행한 결과는 [Table 6]과 같이 나타났다. 기존자료와 새로운 자료의 회귀계수를 비교해 보면 차이가 있는 것을 확인할 수 있었다. LC(간경변), HBV(B형 간염), SEX(남성), α-FP(≥20IU/㎖), Unknown alcohol(미확인 음주력)은 기존 자료보다 새로운 자료에서 회귀계수가 상대적으로 크게 추정되었고, CH(만성 간염), HCV(C형 간염), AGE(연령(≥40세)), ALT(≥40IU/ℓ), Heavy alcohol(상습적 음주자)은 새로운 자료보다 기존 자료의 회귀계수가 상대적으로 크게 추정되었다. ALT(≥40IU/ℓ)같은 경우는 기존 자료에 비해 새로운 자료가 반대의 경향을 보이고 있었다. 위의 결과로 보아 기존 자료를 이용한 예측 모형의 개선이 필요함을 알 수 있었다.

기존 자료를 통해서 얻은 예측 모형에 4가지의 개선방안을 적용시켜 모수를 추정한 결과는 다음과 같다. 개선방안 1은 기존의 예측 모형식을 유지한 방법이고, 개선방안 2는 α의 재보정, 개선방안 3은 α, β의 재보정, 개선방안 4는 모형 축소

를 통한 α와 유의한 β_i들을 재보정한다. 아래의 [Table 7]은 실제로 개선을 위해 추정된 모수를 나타낸 표이다.

Table 7. Apparent parameter of updated versions

Updating method	Parameters estimated	Regression coefficient
Updating method 2	α : intercept	-1.477±0.159
Updating method 3	α : intercept	-1.549±0.241
	$\beta_{overall}$: calibration slope	0.931±0.173
Updating method 4	α : intercept	-1.962±0.277
	$\beta_{overall}$: calibration slope	1.434±0.352
	γ_1 : LC	1.613±0.417
	γ_2 : CH	-1.438±0.414
	γ_3 : HCV	-0.134±0.588
	γ_4 : HBV	0.719±0.566
	γ_5 : AGE(≥40)	-0.230±0.465
	γ_6 : SEX(male)	0.069±0.562
	γ_7 : α-FP(≥20IU/㎖)	-1.818±0.385
	γ_8 : ALT(≥40IU/ℓ)	0.656±0.360
	γ_9 : Heavy alcohol	0.896±0.346
	γ_{10} : Unknown alcohol	1.434±0.352

개선방안 1은 기존 모형을 그대로 사용하여 추가적인 모수 추정이 없었고, 개선방안 2는 α, 개선방안 3은 $\beta_{overall}$, 개선방안 4는 α, $\beta_{overall}$, $\gamma_1,\cdots\gamma_{10}$에 대해 추정하여 나타내었다. 전반적으로 기존 모형의 절편을 줄이는 효과를 보이고 있음을 확인 할 수 있었고, 개선방안 3은 전반적인 계수들의 기울기를 감소시키고, 개선방안 4는 일부 공변량들의 기울기가 증가 및 감소되는 경향을 보이고 있음을 확인할 수 있었다.

[Table 7]에서 추정된 모수를 기반으로 4가지의 개선방안을 적용시켜 얻은 각 독립변수별 회귀계수의 결과는 [Table 8]과 같이 나타난다.

Table 8. Regression coefficient of updated versions

	Regression coefficient			
	Updating method 1	Updating method 2	Updating method 3	Updating method 4
Intercept	-6.254	-7.731	-7.803	-8.216
LC	1.722	1.722	1.603	1.351
CH	0.734	0.734	0.683	0.576
HCV	1.263	1.263	1.176	0.991
HBV	0.775	0.775	0.722	0.609
AGE(≥40)	1.315	1.315	1.224	1.032
SEX(male)	0.300	0.300	0.279	0.235
α-FP(≥20IU/㎖)	0.826	0.826	0.769	0.648
ALT(≥40IU/ℓ)	0.283	0.283	0.263	0.221
Heavy alcohol	0.584	0.584	0.544	0.459
Unknown alcohol	0.222	0.222	0.207	0.175

회귀계수의 추정치는 개선방안 1의 경우는 기존 예측 모형식의 절편과 기울기를 바꾸지 않은 것으로 기존 모형식의 회귀계수와 같다. 개선방안 2는 전반적인 절편을 수정한 것으로 개선방안 1에 비해 절편이 바뀐 것을 확인 할 수 있었다. 새로운 자료를 반영하여 절편이 기존 식에 비해 1.477 더 줄어든 것을 확인 할 수 있었다. 개선방안 3은 전반적으로 절편과 독립변수들의 기울기가 모두 조금씩 바뀐 것을 확인할 수 있었고, 개선방안 1에 비해 줄어든 수치임을 확인 할 수 있었다. 절편이 개선방안 1에 비해 1.549 줄어들고, 독립변수들의 기울기가 개선방안 1의 0.931배로 감소한 것을 확인 할 수 있었다. 개선방안 4는 전체적으로 절편, 기울기 모두 개선방안 1에 비해 가장 많이 줄어 든 것을 확인 할 수 있었다. 절편이 개선방안 1에 비해 1.962 줄고, 각 독립변수들의 기울기가 LC(간경변)는 0.371, CH(만성 간염)는 0.158, HCV(C형 간염)는 0.272, HBV(B형 간염)는 0.166, AGE(≥40세)는 0.283, SEX(남성)는 0.065, α-FP(≥20IU/㎖)는 0.178, ALT(≥40IU/ℓ)는 0.062, Heavy alcohol(상습적 음주자)은 0.125, Unknown alcohol(미확인 음주력)은 0.047 줄어들었다. 전반적인 절편과 기울기를 재보정하고 축소한 방법으로 절편을 포함한 회귀계수가 가장 많이 감소하는 형태를 보였다.

(2) 예측 모형 비교평가

① calibration plot

4.2절에서 구축한 예측모형의 비교 평가를 위해 실제 관측치의 확률값과 개선 모형을 적용시킨 확률값과의 관계를 나타낸 calibration plot을 [Figure 1]에 제시하였다[15,16,17].

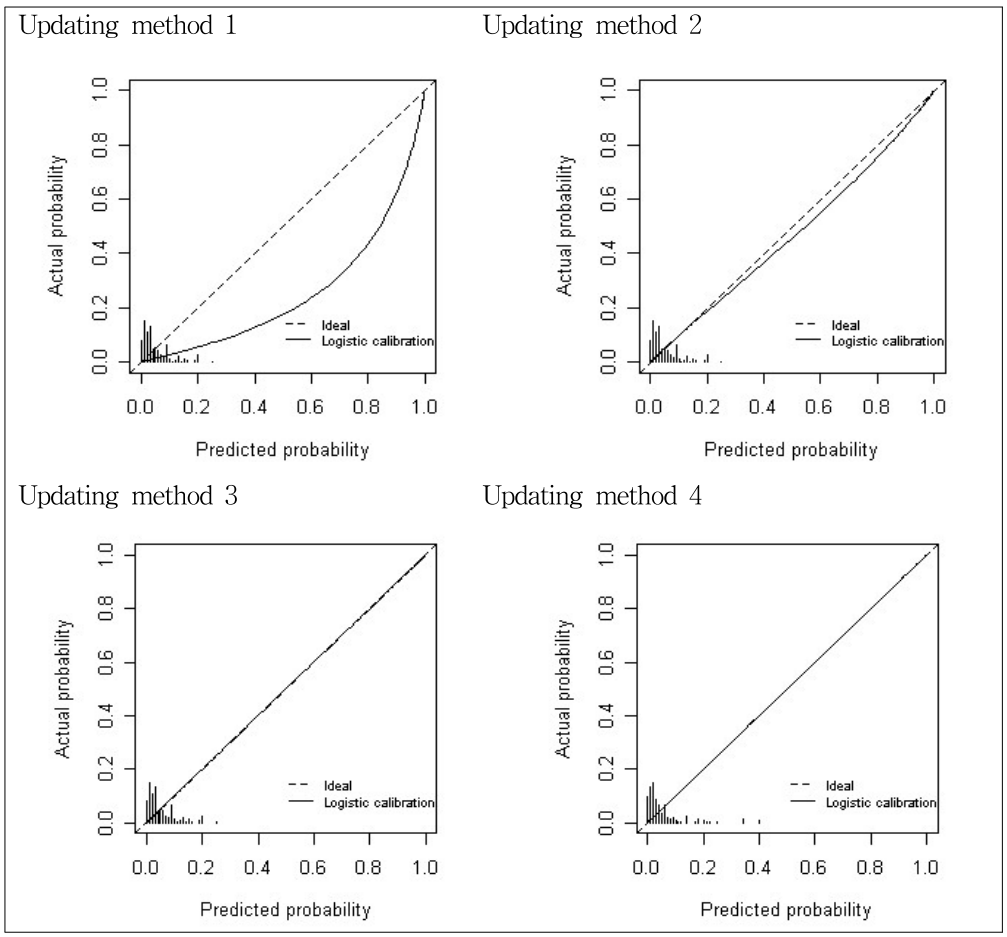

Figure 1. Calibration plot of updated versions

개선방안 1은 기존 모형으로 점선과 실선이 볼록한 형태로 예측 확률과 실제 관측 확률 간의 차이가 크게 나타나고 있고, 개선방안 2는 절편 α만을 재보정 해 줌으로써 차이를 현저하게 줄이고 있다. 그리고 개선방안 3은 거의 점선과 실선이 일치하는 형태를 보이고 있으며, 4의 경우는 점선과 실선이 일치하는 형태를 보이고 있다. 개선방안 중에는 calibration plot으로 비교해 본 결과 개선방안 4가 가장 좋은 방법임을 확인할 수 있었고, 개선방안 2의 경우 예측확률과 실제 관측확률의 차이를 눈에 띄게 감소시켜 전반적인 확률을 조정해 주는 효과가 있음을 확인할 수 있었다. 그리고 완벽한 일치는 아니지만 개선방안 3도 2개의 모수만을 재보정함으로써 점선과 실선을 거의 일치 시켜 효율적인 모형 구축으로의 예측임을 확인할 수 있었다.

② ROC 곡선

모형의 예측력을 나타내는 c 통계량의 비교를 위해 [Figure 2]에서 ROC 곡선을 그려 면적(AUC)를 확인하여 보았다.

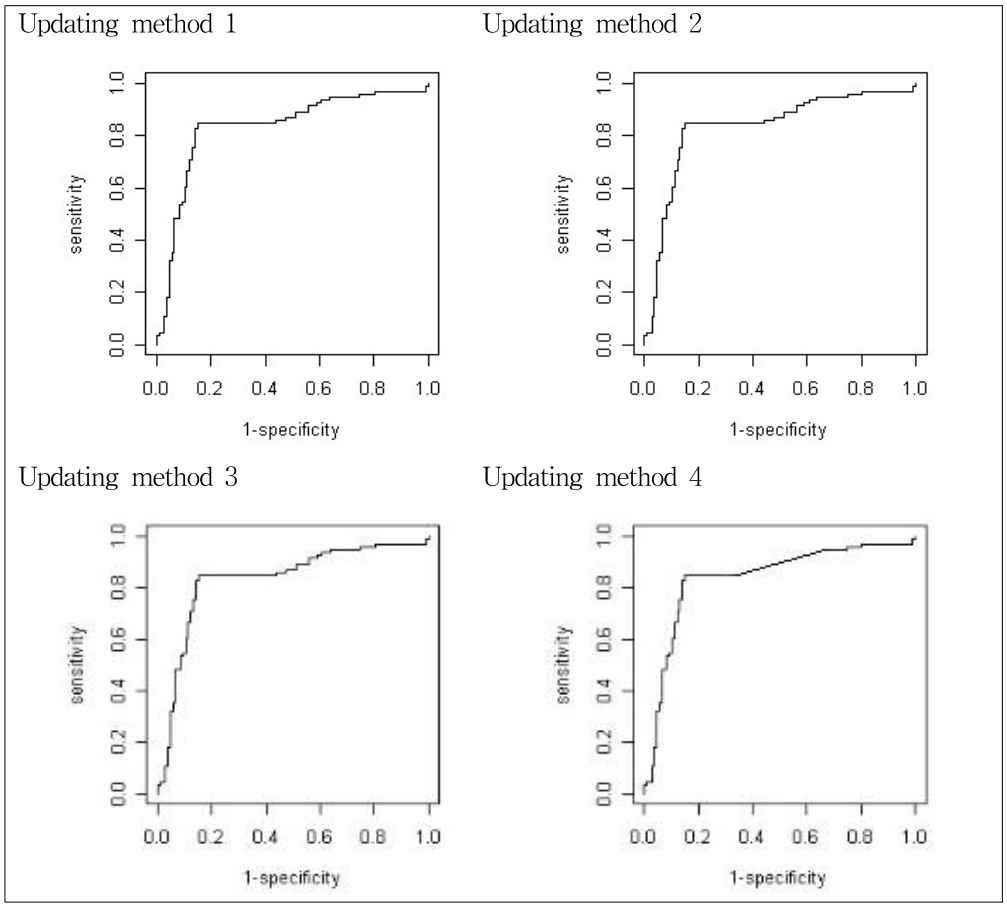

Figure 2. ROC curve of updated versions

개선방안 1, 개선방안 2, 개선방안 3은 ROC 곡선이 똑같이 c 통계량이 0.741로 차이가 없음을 확인 할 수 있었다. 반면에 개선방안 4는 c 통계량이 0.760으로 예측력을 증가시킴을 확인 할 수 있었다. 개선방안 4가 가장 좋은 방법임을 다시 한 번 확인 할 수 있었다.

③ 평가 척도

4가지 개선방안 중 가장 좋은 방안을 선택하기 위하여 [Table 9]에서 평가 척도를 비교하였다.

Table 9. Apparent performance of updated versions

	Updating method 1	Updating method 2	Updating method 3	Updating method 4
Parameters estimated	0	1	2	12
U statistic	0.150	0.000	0.000	0.000
c statistic	0.741	0.741	0.741	0.760
Brier score	0.071	0.048	0.048	0.045
r	0.930	0.930	0.999	1.000

[Table 9]는 모형비교를 위해 나타낸 표로 추정된 모수의 수, 모형이 잘 구축되었는지 여부를 확인하기 위한 보정 평가지수의 척도인 U 통계량, 모형의 예측력을 나타내는 c 통계량, 전반적으로 수행이 잘 되었는지를 확인하기 위한 브라이어 점수, 예측확률과 실제확률의 상관계수 r을 개선방안별로 나타내고 있다. 이 때 U 통계량은 0에 가까울수록, c 통계량은 1에 가까울수록, 브라이어 점수는 0에 가까울수록, r은 1에 가까울수록 잘 개선된 모형으로 위의 표에서는 개선방안 4가 모든 평가 지표에서 가장 좋은 모형임을 확인 할 수 있었다. 반면에 추정된 모수의 수가 많아 다소 복잡한 계산이 따름을 알 수 있었다. 그 밖에 개선방안 2, 개선방안 3도 기존의 식을 개선시킴으로써 개선방안 1에 비해 더 좋은 모형임을 확인 할 수 있었다.

5. 고 찰

지금까지 4가지 개선방안들을 제시하고 비교하였다. 개선방안 1은 기존 자료를 통해 얻어진 예측모형의 절편과 기울기를 바꾸지 않고, 새로운 자료에 적용시킨 방법이다. 4가지 방안 중 가장 접근이 용이한 방법이나, 새로운 자료에 대한 예측력이 상대적으로 떨어지는 것을 확인할 수 있었다. 개선방안 2는 절편을, 개선방안 3은 절편과 기울기를 재보정하여 개선시킴으로써 새로운 자료에 적합한 예측 모형을 구축할 수 있었다. 또한 calibration plot에서 기울기가 거의 1에 가까운 값으로 나타난 것을 확인 할 수 있었다. 개선방안 4는 개별 회귀계수를 재보정과 축소를 이용해 수정함으로써 기존의 자료와 새로운 자료의 특성을 적절히 반영한 예측 모형을 구축할 수 있으며, 상대적으로 영향력이 작은 변수를 제거해주는 특징을 가지고 있었다. 본 연구의 결과 개선방안 4가 가장 좋은 방법으로 나타나고 있었다.

개선방안들은 기존 자료와 새로운 자료의 정보를 모두 반영하고, 많은 자료를 활용하여 예측 모형을 개선함으로써 예측력을 높일 수 있었다.

중요한 점은 기존 모형을 개선하는 방법이므로, 기존 자료에 비해 새로운 자료의 수가 많을 때, 좀 더 유용한 방법이며, 기존에 인자들을 잘 선별할 필요가 있다. 또한 절편 α와 기울기 β를 재보정하는 과정에서 회귀계수의 추정이 정확해야 한다는 것이다. 만약 이 회귀계수가 정확하지 않다면, 잘못된 예측 모형을 유도할 수도 있다.

본 연구에서는 기존 모형에 포함되었던 예측인자만을 이용해서 예측모형을 개선시켰지만, 향후 더 많은 자료를 이용해 분석을 할 경우 기존 모형에 없었던 새로운 위험인자들을 추가하여 예측모형을 개선시키는 방법에 관한 연구가 필요할 것으로 판단된다.

또한 본 연구에서는 로지스틱 회귀모형에서의 개선방법을 다루고 있으나, 일반 회귀분석에서도

적용이 가능한지에 관한 연구를 추후 고려할 수 있을 것이다.

참고문헌

[1] Okuda K, Ohtsuki T, Obata H, et al. Natural history of hepatocellular carcinoma and prognosis in relation to treatment Study of 850 patients. Cancer 1985; 56: 918-928.

[2] Steyberg EW, Borsboom GJJM, Houwelingen HCV, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statistics in Medicine 2004; 23: 2567-86.

[3] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annal of Internal Medicine 1999; 130: 515-24.

[4] 성웅현. 응용 로지스틱 회귀분석, 탐진; 2001.

[5] 정광모, 최용석. 로지스틱 회귀와 응용. 자유아카데미; 2003.

[6] Harrell Jr FE. Regression modeling strategies, Springer 2001.

[7] Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the GUSTO database, Statistics in Medicine 1998; 17: 2501-2508.

[8] Tibshirani R. Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B. 1996; 58: 267-288.

[9] Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of shrinkage techniques in logistic regression analysis: a case study. Statistica Neerlandica 2001; 55: 76-88.

[10] Steyerberg EW, Eijkemans MJC, Harrell Jr FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in Medicine 2000; 19: 1059-1079.

[11] Steyerberg EW, Eijkemans MJC, Houwelingen JCV, Lee KL, Habbema JDF. Prognostic models based on literature and individual patient data in logistic regression analysis. Statistics in Medicine 2000; 19: 141-160.

[12] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning, Springer 2001.

[13] Cheong JY, Han KH, Kim DK, et al. Establishment of Individual Prediction Model According to Risk Factors for Development of Hepatocellular Carcinoma in Korea: Establishment of Individual Prediction Model for Hepatocellular Carcinoma. The Korean Journal of Hepatology 2001; 4: 449-458(Korean).

[14] Choi JW, Ahn SH, Moon CM, et al. Efficacy of Individual Prediction Model for the Early Diagnosis of Hepatocellular Carcinoma. Korean Journal of Medicine 2004; 67: 7-14(Korean).

[15] Harrell Jr FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Statistics in Medicine 1996; 15: 361-387.

[16] Houwelingen HCV. Validation, calibration, revision and combination of prognostic survival models. Statistics in Medicine 2000; 19: 3401-3415.

[17] Steyerberg EW, Vergouwe Y, Keizer HJ, Habbema JDF. Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. Statistics in Medicine 2001; 20: 3847-3859.