연구보고서 2020-20-031

www.nhimc.or.kr

의료이용 행태를 분석하여 중증질환 발생 위험을 예측하는 인공지능 알고리듬의 개발 및 검증: 대규모 한국 코호트 연구

안찬식 · 장정현 · 임현선 · 송선옥 · 노성현 · 최윤정 · 박해용 · 김다인

국민건강보험 National Health Insurance Service
일산병원 Ilsan Hospital

연구보고서
2020-20-031

의료이용 행태를 분석하여 중증질환 발생 위험을 예측하는 인공지능 알고리듬의 개발 및 검증: 대규모 한국 코호트 연구

안찬식 · 장정현 · 임현선 · 송선옥
노성현 · 최윤정 · 박해용 · 김다인

국민건강보험 일산병원 연구소

[저 자]

책임 연구자:	국민건강보험 일산병원 영상의학과	안찬식
공동 연구원:	국민건강보험 일산병원 이비인후과	장정현
	국민건강보험 일산병원 연구소 연구분석부	임현선
	국민건강보험 일산병원 내과	송선옥
	아주대학교 아주대학병원 신경외과	노성현
	연세대학교 용인세브란스병원 병리과	최윤정
	국민건강보험 일산병원 연구소 연구분석부	박해용
	연세대학교	김다인

연구관리번호	IRB 번호
NHIS-2020-2-146	NHIMC 2020-06-033

머리말

간세포암(hepatocellular carcinoma, HCC)은 전세계적으로 암으로 인한 사망 원인 중 3위이며, 대한민국에서는 간세포암을 포함한 간암은 남성에서는 4번째, 여성에서는 6번째로 흔한 암종으로 암 사망 원인 중 2위를 차지하고 있다. 간암 발생건수는 2000년 이후 다소 감소하고 있지만 간암 발생자 수를 보면 오히려 증가 추세를 보이는데, 이러한 결과는 최근 연도로 올수록 고령화가 심화되는 우리나라의 인구구조가 주된 원인일 것으로 생각하고 있다.

최근 하드웨어 발달과 알고리즘 개선으로 딥러닝을 포함하여 기계학습이 다양한 분야에서 적용되고 있으며 많은 발전이 이루어지고 있다. 의학도 예외가 아니며 진단, 치료계획 수립, 예후 예측, 위험군 선별 등 거의 모든 분야에서 인공지능 연구가 활발하게 이루어지고 있다. 국민건강보험 청구자료와 검진자료는 많은 사람들의 많은 정보를 포함하는 빅데이터로서, 인공지능 알고리즘 적용이 고식적인 예측모델보다 우수한 예측력을 가진 모델을 개발하는 데 도움이 될 수 있을 것이라고 생각하였다.

본 연구를 통해 기계학습을 포함한 인공지능 알고리즘으로 국민건강보험공단 청구자료 및 국가검진 결과를 이용하여 검진 수진자의 미래 간세포암 발병 위험을 정확하게 예측할 수 있는 예측모델을 만들고 검증하여 보았다.

끝으로 본 보고서에서 저술한 내용은 저자들의 의견이며, 보고서 내용상의 하자가 있는 경우 저자들의 책임으로 국민건강보험 일산병원 연구소의 공식적인 견해가 아님을 밝혀둔다.

2022년 2월

국민건강보험 일산병원장 김 성 우

일산병원 연구소장 이 천 균

목차

표목차

그림목차

요 약

요 약

1. 연구 배경 및 목적

간세포암(hepatocellular carcinoma, HCC)은 전세계적으로 암으로 인한 사망 원인 중 3위이며, 대한민국에서는 간세포암을 포함한 간암은 남성에서는 4번째, 여성에서는 6번째로 흔한 암종으로 암 사망 원인 중 2위를 차지하고 있다. 간암 발생건수는 2000년 이후 다소 감소하고 있지만 간암 발생자 수를 보면 오히려 증가 추세를 보이는데, 이러한 결과는 최근 연도로 올수록 고령화가 심화되는 우리나라의 인구구조가 주된 원인일 것으로 생각하고 있다. 2005년 암으로 인한 사회경제적 부담을 각 암종별로 나누어 살펴보았을 때, 간암의 부담 수준이 2조3,963억원의 위암을 추월하고 1위를 차지하였다. 따라서, 발생률이 감소 추세로 들어서기는 했으나, 간세포암종은 여전히 한국인에게 큰 부담을 주는 질병임을 확인할 수 있다.

간암은 다른 대부분의 암과 달리 발생 가능성이 높은 고위험군이 뚜렷하게 알려져 있다. B형간염, C형간염, 알코올 간질환 등의 상당수에서 간경변증을 거쳐 간암이 발생한다. 따라서 간경변증이 있거나 B형간염바이러스 또는 C형간염바이러스의 보유자는 간암의 고위험군으로 감시검사의 대상이 된다. 그러나 본인이 간암 고위험군이라는 것을 인지하지 못하고 있는 경우가 많고, 인지하더라도 감시검사를 성실하게 받지 않는 경우가 있을 수 있다.

거의 모든 대한민국 국민은 의무적으로 국민건강보험에 가입되어 있거나 의료보호의 대상이 되며, 40세 이상의 성인은 최소 2년에 한 번 국가가 제공하는 건강검진을 받도록 되어있다. 이러한 보험청구 내역과 검진 결과는 데이터베이스화 되고 있다. 청구내역과 건강검진결과에는 간암의 위험인자들, 예를 들어 나이, 성별 등의 사회인구학적 정보, 만성간염바이러스질환 등의 과거력, 간효소검사결과 등에 대한 정보가 포함되어 있다. 따라서, 현재 행해지고 있는 국가검진이 간암 선별 혹은 감시검사의 목적은 갖고 있지 않더라도, 이러한 정보를 이용하여 간암 발생의 위험이 높은 사람에게 경고를 해줄 수 있다면 데이터로부터 부가적인 가치를 생성하는, 즉 잠재적 간세포암 고위험 환자를

발견하여 질병 예방 혹은 조기진단의 기회를 줄 수 있는 의미 있는 일이 될 것이라 생각하였다.

최근 하드웨어 발달과 알고리즘 개선으로 딥러닝을 포함하여 기계학습이 다양한 분야에서 적용되고 있으며 많은 발전이 이루어지고 있다. 의학도 예외가 아니며 진단, 치료계획 수립, 예후 예측, 위험군 선별 등 거의 모든 분야에서 인공지능 연구가 활발하게 이루어지고 있다. 국민건강보험 청구자료와 검진자료는 많은 사람들의 많은 정보를 포함하는 빅데이터로서, 인공지능 알고리즘 적용이 고식적인 예측모델보다 우수한 예측력을 가진 모델을 개발하는 데 도움이 될 수 있을 것이라고 생각하였다.

따라서 본 연구의 목적은 국민건강보험공단 청구자료 및 국가검진 결과에 근거하여 검진 수진자의 미래 간세포암 발병위험도를 예측하는 기계학습 알고리즘의 개발 및 검증이었다.

2. 연구 결과

최종 연구대상자는 41,7346명이었고 79.5%인 33,1694명이 학습군, 나머지 20.5%인 8,5652명이 테스트군에 속하였다. 건강검진 수진일 기준으로 전체 대상자의 평균 나이는 55세였고 그 범위는 42세에서 82세까지였다. 남성 대 여성의 비는 약 5.5:4.5로 남성이 좀 더 많았다.

추적관찰기간의 중간값은 학습군에서 11.1년(최대 12.0년까지)이었고 테스트군에서 9.1년(최대 10.0년까지)였다. 학습군 33,1694명 중 0.5%인 1799명과 테스트군 8,5652명 중 0.5%인 390명이 관찰기간 중 간세포암이 발병하였다. 간세포암 외 다른 암종은 학습군 33,1694명 중 8.4%인 27856명, 테스트군 8,5652명 중 7.9%인 6732명에서 관찰기간 중 발생하였다.

1000번의 부트스트랩 표본에서 85% 이상 간세포암 발병과 유의한 관계를 보였던 변수들은 나이, 성별, 비만, 소득수준, 만성간질환의 가족력, aspartate transaminase (AST), alanine transaminase (ALT), gamma-glutamyl transferase (GGT), 혈중 총콜레스테롤, 만성간염바이러스 감염력, 인체면역결핍바이러스 감염력, 당뇨병, 고지혈증, 그리고 조현병 등 정신질환이었다.

다변수 콕스 비례-위험 회귀분석에서, 더 나이가 많을수록(hazard ratio [HR], 1.581 / 10살), 여성보다 남성에서(HR, 3.122), 만성간질환의 가족력이 있을수록(HR, 2.490),

비만인 경우(HR, 1.648), ALT 수치가 높은 경우(HR, 1.049 / 10 IU/L), GGT 수치가 높은 경우(HR, 1.030 / 10 IU/L), 만성간질환이 있는 경우(HR, 3.430), 만성간염바이러스감염력이 있는 경우(HR, 1.851), HIV 감염력 있는 경우(HR, 4.097), 당뇨병이 있는 경우(HR, 1.427)에 간세포암 발병위험이 더 높았다. 반면, 혈충 총 콜레스테롤이 높은 경우(HR, 0.897 / 10 mg/dL), 고지혈증이 있는 경우(HR, 0.479), 조현병 등 정신질환이 있는 경우(HR, 0.655), 소득수준이 높은 경우(HR, 0.832)는 간세포암 발병위험이 더 낮았다(모든 경우 $p < 0.001$). 간세포암 외 다른 암종을 경쟁인자로 두고 분석하였을 때도 비슷한 결과를 얻었다.

학습군에서, 간세포암 발병 위험을 예측하는데 있어 XGBoost 모델이 생존랜덤포레스트 모델보다 더 우수한 성능을 보였다. 간세포암이 관찰기간 내 발병할지 아니면 하지 않을지 이분법적으로 판단하는데 있어서, area under the cureve (AUC) (+/- 표준편차)는 XGBoost와 생존랜덤포레스트가 교차검증에서 각각 0.882 (+/-0.013)과 0.871 (+/-0.019)이었다. Calibration 도표에서 두 알고리즘은 비슷하게 좋은 성능을 나타내었다. Brier skill score는 각각 0.109와 0.062로 이는 모든 대상자에서 간세포암이 발병하지 않을 것이라 예측한 경우와 비교하여 XGBoost와 생존랜덤포레스트 모델을 사용하였을 때 각각 10.9%와 6.2%의 Brier score 향상이 예상된다는 의미이다. 두 알고리즘의 앙상블 모형이 가장 좋은 성능을 보였는데, AUC, Brier skill score는 각각 0.892 (+/-0.011)과 0.112였다. 따라서 앙상블 모델이 우리의 최종모델로 선택되었다.

최종적으로 학습된 예측모델을 테스트군에서 최종 검증하였다. 20% 이하 확률에서 다소 위험을 과소평가하는 경향을 보였으나 전반적으로 좋은 calibration을 보였다. AUC는 0.873이었고 95% 신뢰구간은 0.860-0.885였다. Brier skill score는 0.078로 기저평가 기준에 비해 약 7.8% 정도 Brier score의 향상이 있었다. 발병확률 1%을 기준으로 그 이상인 경우 발병 위험이 있다고 보았을 때의 진단적 민감도, 특이도, 정확도는 각각 71.8% (95% 신뢰구간, 71.4-72.2), 88.4% (95% 신뢰구간, 88.1-88.7), 88.4% (95% 신뢰구간, 88.2-88.6)이었다.

간세포암 발병까지 걸린 시간의 중간값은 학습군에서 약 294주 혹은 5.6년, 테스트군에서 약 235주 혹은 4.5년이었다. 발병까지 걸린 시간을 예측하는 데 있어서 생존랜덤포레스트 모델이 콕스 비례-위험 모델보다 우수한 성능을 보였다. 테스트군에서, 콕스 모델은 AUC가 0.828 (95% 신뢰구간 0.819-0.838)이었던 반면 생존랜덤포레스트 모델은 0.857 (95% 신뢰구간 0.850-0864)로 95% 신뢰구간이 겹치지 않아 유의한 차이가 있는

것으로 해석되었다.

3. 결론 및 제언

본 연구를 통해 국가검진 후 약 10년 내 간세포암이 발병할 위험을 검진결과와 국민건강보험 청구자료에 근거하여 예측하는 기계학습 모델을 구축하고 검증하였다. 이 예측모델은 검증 결과 학습 시 예상했던 성능을 보임을 확인하였다. 더 나아가, 당뇨병 및 지방간이 있는 특정 질환군에서도 같은 결과를 얻었다. 이전에 출판된 모델들은 대부분 만성간질환이 이미 있어 간세포암 고위험군으로 분류되는 환자들을 대상으로 하고 Cox 비례-위험 회귀분석을 이용한 모델이었다. 반면 본 연구의 모델은 고위험군은 물론 전통적인 간암 위험인자가 없는 사람들까지 모두 대상으로 하면서, 기계학습 알고리즘인 랜덤포레스트와 XGBoost 알고리즘을 이용하여 좋은 성능을 확인하였다.

본 연구진은 이번 연구를 통하여 기계학습을 포함한 인공지능 알고리즘으로 국민건강보험공단 청구자료 및 국가검진 결과를 이용하여 검진 수진자의 미래 간세포암 발병 위험을 정확하게 예측할 수 있는 예측모델을 만들고 검증하여 보았다. 추후 이러한 노력들이 실제 적용되어, 검진 수진자들이 일반적인 건강상태 정보에 더하여 암과 같은 중요한 질환의 발병 위험과 그 원인이 되는 인자를 보고받고 이에 따라 병 예방을 위해 노력할 수 있도록 돕는 시스템이 갖추어지길 희망한다.

제 1 장

서론

제1장 서론

제1절 연구 배경 및 필요성

간세포암(hepatocellular carcinoma, HCC)은 전세계적으로 암으로 인한 사망 원인 중 3위이며 발병률은 매 해 50만명을 넘는다.[1,2] 대한민국에서는 간세포암을 포함한 간암은 남성에서는 4번째, 여성에서는 6번째로 흔한 암종이며, 암 사망 원인 중 2위를 차지하고 있다.[3] 간암 발생건수는 2000년 이후 다소 감소하고 있는데, 이는 국내 간암의 주원인이라 할 수 있는 B형간염이 예방접종사업의 영향으로 감소하고 있는 것에 기인하고 있는 것으로 추정된다. 한편, 간암 발생자 수를 보면 오히려 증가 추세를 보이는데, 이러한 결과는 최근 연도로 올수록 고령화가 심화되는 우리나라의 인구구조가 주된 원인일 것으로 생각하고 있다. 또한, 5년 유병자 수를 기준으로 보면, 2007년 간암의 연령표준화 유병률은 43.0명이었으며, 2008년 45.1명, 2009년 46.1명, 2010년 46.6명으로 점차 증가하는 양상을 보인다. 간암 발생률은 감소하는데 유병률이 증가하는 양상을 보이는 것은 전반적으로 간암의 생존율이 증가하기 때문인 것으로 추정된다. 2005년 암으로 인한 사회경제적 부담을 각 암종별로 나누어 살펴보았을 때, 간암의 부담 수준이 2조 3,963억원의 위암을 추월하고 1위를 차지하였다. 따라서, 발생률이 감소 추세로 들어서기는 했으나, 간세포암종은 여전히 한국인에게 큰 부담을 주는 질병임을 확인할 수 있다.

간암은 다른 대부분의 암과 달리 발생 가능성이 높은 고위험군이 뚜렷하게 알려져 있다. B형간염, C형간염, 알코올 간질환 등의 상당수에서 간경변증을 거쳐 간암이 발생한다. 따라서 간경변증이 있거나 B형간염바이러스 또는 C형간염바이러스의 보유자는 간암의 고위험군으로 감시검사의 대상이 된다. 그러나 본인이 간암 고위험군이라는 것을 인지하지 못하고 있는 경우가 많고, 인지하더라도 감시검사를 성실하게 받지 않는 경우가 있을 수 있다.

거의 모든 대한민국 국민은 의무적으로 국민건강보험에 가입되어 있거나 의료보호의 대상이 되며, 40세 이상의 성인은 최소 2년에 한 번 국가가 제공하는 건강검진을 받도록

되어있다. 이러한 보험청구 내역과 검진 결과는 데이터베이스화 되고 있다. 청구내역과 건강검진결과에는 간암의 위험인자들, 예를 들어 나이, 성별 등의 사회인구학적 정보, 만성간염바이러스질환 등의 과거력, 간효소검사결과 등에 대한 정보가 포함되어 있다.[4] 따라서, 현재 행해지고 있는 국가검진이 간암 선별 혹은 감시검사의 목적은 갖고 있지 않더라도, 이러한 정보를 이용하여 간암 발생의 위험이 높은 사람에게 경고를 해줄 수 있다면, 데이터로부터 부가적인 가치를 생성하는, 즉 잠재적 간세포암 고위험 환자를 발견하여 질병 예방 혹은 조기진단의 기회를 줄 수 있는 의미 있는 일이 될 것이라 생각하였다.

제2절 연구 목적

지금까지 간세포암 발생 위험을 예측하는 여러 예측모델들이 발표되었다.[5-12] 그러나 대부분의 이전 모델들은 이미 고위험군으로 확인된 사람들에게만 적용하도록 되어 있으며 고위험군이 아닌 사람들은 대상이 되지 않는다. 따라서 고위험군에 속하지 않거나 그 여부를 모르는 사람들까지 포함하여 모든 사람을 대상으로 간세포함 발생위험을 예측할 수 있는 모델은 기존에 식별되지 않았던 간세포암 고위험군을 선별하는데 부가적인 중요한 역할을 할 수 있을 것이다. 거의 모든 국민의 의료사용내역과 일반건강검진결과가 반강제적으로 얻어지고 데이터베이스화 되므로, 이 자료를 이용하여 간세포암 발생위험을 예측하는 것은 이러한 목적에 정확히 부합한다.

최근 하드웨어 발달과 알고리즘 개선으로 딥러닝을 포함하여 기계학습이 다양한 분야에서 적용되고 있으며 많은 발전이 이루어지고 있다. 의학도 예외가 아니며 진단, 치료계획 수립, 예후 예측, 위험군 선별 등 거의 모든 분야에서 인공지능 연구가 활발하게 이루어지고 있다. 국민건강보험 청구자료와 검진자료는 많은 사람들의 많은 정보를 포함하는 빅데이터로서, 인공지능 알고리즘 적용이 고식적인 예측모델보다 우수한 예측력을 가진 모델을 개발하는 데 도움이 될 수 있을 것이라고 생각하였다.

따라서, 본 연구에서는 청구자료와 건강검진결과를 바탕으로 건강검진 수진자의 향후 간세포암 발생위험을 인공지능 알고리즘으로 예측하는 모델을 개발하고 검증고자 하였다. 모델 개발에는 국민건강보험공단에서 연구용으로 구축하여 공개한 건강검진 표본코호트(National Health Insurance Service-National Health Screening, NHIS-HEALS)를 이용하였다.[13,14]

제2장

이론적 고찰

제2장 이론적 고찰

제1절 기계학습 알고리즘

1. 콕스 비례-위험 모형(Cox proportional-hazard model)

콕스 회귀모형, 혹은 콕스 비례-위험 모형이란 생존과 관련된 여러 설명변수가 있을 때 생존에 영향을 미치는 여러 변수들을 동시에 알아보기 위해 사용되는 분석법이다. 생존시간, 즉 본 연구에서는 건강검진 이후 간세포암이 발생하기 전 까지의 시간에 대해 어떠한 분포형태도 가정하지 않으므로 비 모수적인 분석이지만, 모형에 근거하여 회귀계수를 추정한다는 점이 모수적방법과 유사하여 준모수모형 이라고 한다. 또한, 콕스 회귀모형은 t시점에서의 로그(log) 위험함수를 여러 설명변수 들의 선형식으로 표현한다.

만약 p개의 설명변수가 있는 콕스 모형에서 i번째 건강검진 수진자의 설명변수 값이 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$이고, 회귀 모형계수가 $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$이라면 콕스 모형은 다음 [그림 2-1]에 나타낸 식으로 표현된다.

$$h_i(t) = h_0(t)\exp(\beta' x_i)$$
$$= h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$$

[그림 2-1] 콕스 비례-위험 모형 식

여기서, $h_0(t)$는 기저위험함수를 의미하며, 위험함수에 미치는 여러 설명변수들의 영향이 전혀 없을 경우를 가정한다. 하지만 실제적으로 모든 수치가 0값을 가질 때는 불가능하므로 각 변수를 전체 수진자의 평균값으로 뺀 변수로 다시 정의한다. 또한, 위험함수(hazard function), 즉 h(t)는 [그림 2-2]에 나타낸 식과 같으며 t시점까지 생존한 기업이 t시점 바로 직후에 간세포암이 발병할 조건부 확률로 순간위험률이라고 하며, S(t)는 생존함수로 t시점까지 발병하지 않을 확률이고, f(t)는 t시점에 발병할 확률을 의미한다.

$$h(t) = \lim_{\triangle t \to 0} \frac{\Pr(t \le T < t + \triangle t \mid T \ge t)}{\triangle t}$$
$$= \frac{1}{\Pr(T \ge t)} \lim_{\triangle t \to 0} \frac{\Pr(t \le T < t + \triangle t}{\triangle t} = \frac{f(t)}{S(t)}$$

[그림 2-2] 위험함수 식

콕스 회귀모형은 시간에 관계없이 위험이 비례함수를 따른다는 가정을 하며, 이것이 콕스 "비례" 위험회귀모형이라고 불리는 이유다. 따라서, 콕스 회귀모형은 변수들이 비례성 가정을 만족하는지 여부를 판단하여야 한다. 비례성 가정에 대한 가장 쉬운 판단 방법은 LLS 도식을 그려 시각적으로 검토하는 방법이 있으며, 좀 더 객관적인 방법으로 시간 의존적인 변수에 대한 검정통계량을 이용하여 판단할 수도 있다.

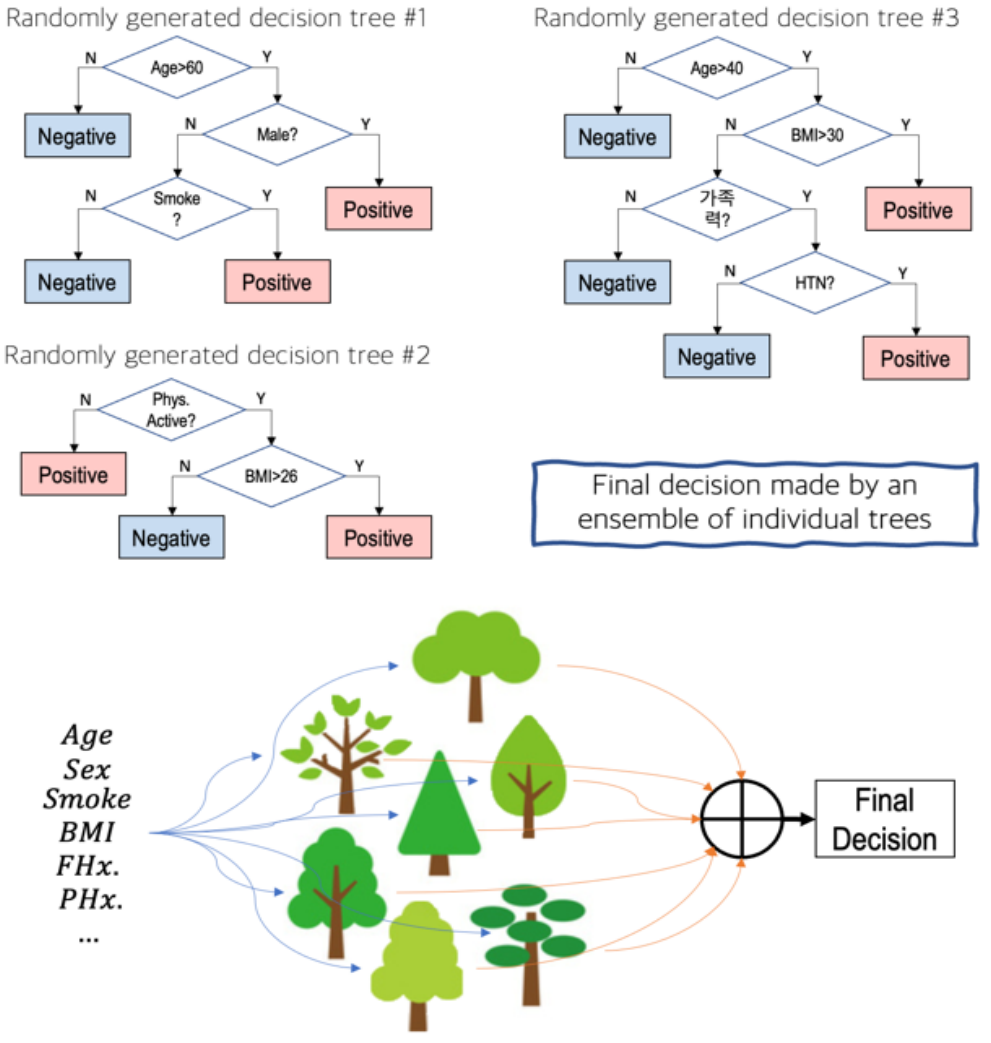
2. 랜덤포레스트(random forest) 알고리즘

기계 학습에서의 랜덤포레스트(random forest)는 분류, 회귀, 생존분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정나무(decision tree)로부터 분류 또는 평균 예측치(회귀 분석)를 얻는 방법이다. 앙상블 학습이란 여러 개의 모델을 사용해서 각각의 예측 결과를 만들고 각 결과를 종합하여 결론을 내리는 것으로, 일종의 "집단지성"의 힘을 빌리는 것과 유사하다. 랜덤포레스트에서의 앙상블은 무작위로 생성된 결정나무에서 추출된 결과들을 종합하여 최종 결과를 얻는 것을 의미한다.

결정나무(decision tree)는 말 그대로 결정을 내리기 위해 사용하는 나무 형태의 구조를 갖는 알고리즘으로, 결정 과정을 간단한 문제들로 이루어진 계층 구조로 나눈다. 간단한 문제에 대해서는 매개변수(예: 모든 노드의 테스트 매개변수, 종단 노드에서 매개변수 등)를 사용자가 직접 설정할 수 있지만, 보다 복잡한 문제의 경우 학습 데이터로부터 나무 구조와 매개변수를 모두 자동으로 학습한다. 일반적으로 결정나무를 이용한 방법의 경우, 그 결과 또는 성능의 변동 폭이 크다는 결점을 가지고 있다. 특히 학습 데이터에 따라 생성되는 결정나무가 무작위성(randomness)에 따라 매우 다르기 때문에 일반화하여 사용하기에 어려움이 따른다. 특히, 결정나무는 계층적 접근방식이기 때문에 만약 중간에 오류가 발생한다면 다음 단계로 오류가 계속 전파되는 특성을 가진다.

랜덤포레스트 알고리즘은 이러한 결정나무의 단점을 많은 수의 서로 다른 결정나무를 만들어 각 결정나무의 결정을 종합하여 결정을 내려서 극복한다. 만들어지는 결정나무의 깊이 및 사용하는 설명변수의 수와 종류 모두 무작위로 정함으로써 나무들이 서로 조금씩 다른 특성을 갖게 한다. 이로써 각 트리들의 예측 결과는 해결해야 할 과제를 다양한

관점에서 해석한 것이 되어 소위 비상관화(decorrelation)되며, 결과적으로 예측모델의 일반화(generalization) 성능을 향상시킨다. 또한, 이러한 무작위성이 예측모델이 잡음, 즉 의미 없는 정보가 포함된 데이터에 대해서도 강인하게 만들어 준다. 무작위성은 각 나무들의 훈련 과정에서 발생되며, 무작위 학습 데이터 추출 방법을 이용한 앙상블 학습법인 배깅(bagging)과 무작위 노드 최적화(randomized node optimization)가 자주 사용된다. 이 두 가지 방법은 서로 동시에 사용되어 무작위성을 더욱 증진 시킬 수 있다.



[그림 2-3] 랜덤포레스트 알고리즘

배깅(bagging)은 bootstrap aggregating의 약자로, 부트스트랩(bootstrap)을 통해 조금씩 다른 훈련 데이터에 대해 훈련된 기초분류기(base learner)들을 결합(aggregating)시키는 방법이다. 부트스트랩이란, 주어진 훈련 데이터에서 중복을 허용하여 원 데이터셋과 같은 크기의 데이터셋을 만드는 과정을 말한다. 결정나무는 작은 편향과 큰 분산을 갖기 때문에, 매우 깊이 성장한 나무는 훈련 데이터에 대해 과적합하게 된다. 부트스트랩 과정은 결정 트리들의 편향은 그대로 유지하면서, 분산은 감소시키기 때문에 랜덤포레스트의 성능을 향상시킨다. 즉, 한 개의 결정 나무의 경우 훈련 데이터에 있는 잡음에 대해서 매우 민감하지만, 결정나무들이 서로 상관화되어 있지 않다면 여러 나무들의 평균은 잡음에 대해 강인해진다. 랜덤포레스트를 구성하는 모든 트리들을 동일한 데이터셋으로만 훈련시키게 되면, 트리들의 상관성은 굉장히 커질 것이다. 따라서 배깅은 서로 다른 데이터셋들에 대해 훈련 시킴으로써, 트리들을 비상관화시켜 주는 과정이다.

3. 랜덤포레스트 생존 예측(random survival forest) 알고리즘

랜덤 포레스트 생존 예측(Random survival forest) 알고리즘은 생존 자료 분석에 적용하기 위해 랜덤포레스트 알고리즘을 발전시킨 것이다(Ishwaran 외 2008). 모델 수립은 대략적으로 다음과 같은 과정 거치게 된다.

먼저, 주어진 데이터셋으로 B개의 부트스트랩 표본을 생성한다. 선택되지 않은 데이터는 out-of-bag 표본으로 두고, 나머지 in-bag 표본으로 생존 나무를 성장시킨다. 각 마디에서 p개의 후보변수를 무작위를 골라, 이 중에서 자식 마디의 동질성이 최대가 되는 변수를 선택하여 최적의 분리가 발생되는 지점을 찾는다. 미리 정해진 기준에 도달할 때까지 이 과정을 반복하며 마디를 분리해 나간다. 마디 간 차이는 생존 시간이 다름을 의미하므로, 분리 규칙은 로그 랭크(log rank) 검정 통계량이 최대가 되는 변수와 지점을 찾아 생존 차이를 극대화하는 것이다. 마디의 불순도(impurity)는 생존 여부와 시간과 관련이 있으므로 자료에는 생존 시간과 우중도 절단 여부에 대한 정보가 반드시 포함되어야 한다. 마디가 더 이상 분리되지 않은 지점에 도달하면, 그 마디를 끝마디라 한다.

다음 단계는 모든 B개의 생존 트리의 끝마디에서 얻은 정보를 결합하여 앙상블 누적위험함수(ensemble cumulative hazard function)을 추정하는 것이다. B개 중 b번째 in-bag 부트스트랩 표본에서 형성된 생존 트리의 h번째 끝마디에 N(h)번의 사망 사건이 포함되어 있다면, 관측된 시점을 $T_{1,h} < T_{2,h} < \cdots < T_{N(h),h}$ 으로 표현한다. 이 때 $d_{s,h}$를 시점 s까지 사망이 발생한 횟수, $Y_{s,h}$를 시점 $t_{s,h}$에서 위험 상태에 있는 개체 수라 한다면, h번째 마디의 누적위험함수는 Nelson-Aalen의 추정치로 [그림 2-4]의 식 (1)과 같고,

동일한 마디에 포함된 개체들은 모두 같은 누적위험함수를 따른다.

$$(1)\quad \hat{H}_h(t) = \sum_{t_{s,h} \le t} \frac{d_{s,h}}{\tilde{Y}_{s,h}}$$

$$(2)\quad H(t|\boldsymbol{X}_i) = \hat{H}_h(t)\,, \qquad \boldsymbol{X}_i \in h$$

$$(3)\quad \hat{H}^*_e(t|\boldsymbol{X}_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{H}^*_b(t|\boldsymbol{X}_i)$$

$$(4)\quad \hat{S}^*(t|\boldsymbol{X}_i) = \exp(-\frac{1}{B}\sum_{b=1}^{B} \hat{H}^*_b(t|\boldsymbol{X}_i))$$

[그림 2-4] 생존 랜덤포레스트 알고리즘 식

모형 구축에 사용된 모든 개체는 반드시 하나의 끝마디에 속하고, 개체 i는 p차원의 공변량 벡터 X_i로 설명된다. 따라서 벡터 X_i는 반드시 1개의 끝마디에 속할 수 있으므로, 개체 i의 누적 위험함수는 [그림 2-4]의 식 (2)와 같이 정의할 수 있다. In-bag 부트스트랩 표본을 이용한 앙상블 누적위험함수도 마찬가지로 Nelson-Aalen 추정량에 근거한다. 여기서 b번째 in-bag 부트스트랩 표본에서 생성한 트리의 누적위험함수를 $H_b^*(t|X)$라 하면, 식 (3)과 같이 i의 앙상블 누적위험함수는 B개의 나무를 결합한 평균이다. 그리고 앙상블 생존함수는 위험함수와 생존함수의 관계에 따라 식 (4)와 같이 표현한다.

4. XGBoost (extreme gradient boosting)

XGBoost는 그래디언트 부스팅(gradient boosting)의 한 종류이나, 그래디언트 부스팅의 단점인 느린 수행시간 및 과적합 규제 부재 등의 문제를 해결하기 때문에 결정나무 기반의 앙상블 학습에서 가장 많은 주목을 받은 알고리즘 중 하나이다. 최근 각종 실전 과제에서 예측 또는 분류에 있어서 일반적으로 다른 기계학습 알고리즘보다 뛰어난 성능을 나타내어 많은 인기를 얻고 있다.

XGBoost의 기본적인 개념은 약한 결정나무(week decision tree)를 묶어서 강한 분류기(strong classifier)를 만드는 것으로 랜덤포레스트와 비슷하지만, XGBoost의 경우에는 추가하는 결정나무를 정하는 데 규칙을 둔다는 차이가 있다. 따라서 "boosting"이라고 명명한 것이다. 결정나무를 순차적으로 늘려가는 과정에서, 이전 결정나무에서 얻은 정보를 다음 결정나무를 생성하는 데 활용하는 과정을 통해 이전의 오차를 교정해 나가

는 것이다. 일반적인 그래디언트 부스팅 방식에서는 가지를 쳐나가는 과정에서 오류가 커지면 그 과정을 바로 중단하는 반면, XGBoost는 모델 적합 시 지정된 파라미터에 따라 계속 가지치기(tree pruning)를 수행한 후 개선이 일정 수준에 못 미칠 경우 역방향으로 가지치기를 진행한다. XGBoost는 결측치를 내부적으로 처리하며 결정나무를 생성할 때 병렬적으로 처리하는데, 이러한 과정을 컴퓨터의 중앙처리장치가 다수의 코어를 가지고 있는 경우 동시에 병렬로 작업을 수행할 수 있기 때문에 훨씬 빠르게 분석이 가능하다.

제2절 예측모델의 평가지표

1. Concordance index (C-index)

생존분석에서 가장 많이 사용하는 평가지표이다. 본 연구와 같이 이분형 종속변수를 갖는 경우는 area under the curve (AUC)와 같다. 대상의 정확한 생존 시간을 평가하지는 않고 여러 대상의 생존 시간(또는 위험)을 상대적으로 비교하여 사망 순서를 잘 예측하는지 판단하는 지표로, 모델의 분류력(discriminative ability)를 평가한다. [그림 2-5]의 식은 대상 한 쌍을 비교할 때의 일치확률(concordance probability)를 정의한 것으로, y는 사건이 발생한 실제 시각이며 $\hat{y}$은 모델이 예측한 시간이다.

$$c = \Pr\left(\widehat{y_1} > \widehat{y_2} \mid y_1 \geq y_2\right)$$

[그림 2-5] Concordance 확률 식

이 정의를 바탕으로 concordance index (C-index)를 [그림 2-6]과 같이 계산할 수 있다.

$$\hat{c} = \frac{1}{P'} \sum_{i:\,\delta_i = 1} \sum_{j:\,y_i < y_j} I\left[S\left(\widehat{y_i} \mid X_i\right) < S\left(\widehat{y_i} \mid X_j\right)\right]$$

[그림 2-6] C-index 계산식

P′는 평가 대상 쌍 개수이며, I는 주어진 조건이 참인 경우를 세는 indicator 함수이다. 전체 평가 대상 쌍 중 대상 i보다 오래 생존한 대상 j의 생존함수를 더 크게 예측한 쌍의 비율을 계산하며, 0과 1 사이의 값을 가진다. 집계 조건 중 $\delta_i = 1$은 대상 i에 반드시

사건(즉, 발병)이 발생해야 한다는 의미이다. 반대로 대상 i에 사건이 발생하기 전에 중도절단(censored)됐다면 대상 j가 대상 i보다 오래 생존했다고 확신할 수 없기 때문에 비교 대상에서 제외한다. P′는 결국 전체 비교 가능 쌍 중 대상 i에 사건이 발생한 쌍의 개수이다.

2. Calibration

예측모형에서 calibration이란 예측된 모형이 얼마나 잘 적합한가를 알아보는 척도이다. Calibration은 사건이 발생할 확률의 크기를 이용하여 M개의 집단으로 분할하고, 각 집단에서 그 집단에 속한 개인들의 평균 예측 확률과 실제 결과를 비교하여 평가한다. 로지스틱 회귀분석에서 흔히 쓰이는 것이 Hosmer-Lemeshow 통계량인데, [그림 2-7]의 식과 같다.

$$\chi^2 = \sum_{j=1}^{M} \frac{\left[\frac{Y_j}{n_j} - \overline{P_j}\right]^2}{\overline{P_j}(1-\overline{P_j})}, \qquad j = 1, 2, \cdots, M$$

[그림 2-7] Hosmer-Lemeshow 통계량

중도절단된 자료를 가진 생존모형에서 적합도를 평가하기 위해 Nam과 D'Agostino가 제안한 통계량은 [그림 2-8]의 식과 같다.

$$\chi^2 = \sum_{j=1}^{M} \frac{n_j[KM_j - \overline{P_j}]^2}{\overline{P_j}(1-\overline{P_j})}, \qquad j = 1, 2, \cdots, M$$

[그림 2-8] Nam-D'Agostino 통계량

여기서 KM_j과 P_j는 j번째 Kaplan-Meier 추정치, 그리고 Cox 비례-위험 보형으로 추정한 사건 발생 확률의 평균이다.

3. Brier skill score

Brier score는 전체적인 정확도를 비교하는 계산식으로 관찰된 결과와 예측된 결과와의 차이의 제곱의 평균값으로 0점부터 0.25점 사이로 계산된다. [그림 2-9]의 계산식에서, f는 예측한 결과이고 o는 실제 결과이다. 0점은 완벽한 예측 모델이고 0.25는 무의미

한 예측모델을 나타낸다.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

[그림 2-9] Brier score 식

하지만 Brier score는 발생 확률이 낮은 경우에는 단순히 모든 경우에 발병이 없을 것이라고 하여도 높은 정확도(다시 말해 0에 가까운 Brier score)가 나올 수 있다는 단점이 있다. 본 연구도 많은 검진 수진자 중 극히 일부에서만 간암이 발병하기 때문에 C-index가 매우 낮아도 Brier score는 매우 높게 나올 수 있다.

이러한 점을 극복하기 위해 기저평가(baseline evaluation) 점수와 비교하여 상대적으로 얼마나 Brier score가 향상되었는지 평가하는 Brier skill score를 사용하며, [그림 2-10]의 식과 같이 계산된다. 이 식에서 BS는 예측모형의 Brier score를, BS_{ref}는 기저평가 시 Brier score 이다. 본 연구에서는 기저평가는 모든 경우 간암이 발병하지 않을 것이라 예측한 경우로 하였다.

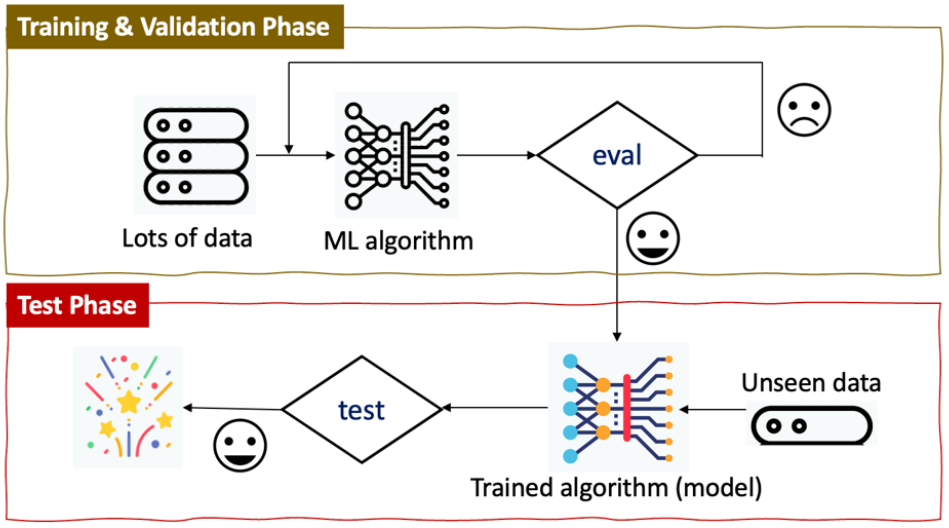
$$BSS = 1 - \frac{BS}{BS_{ref}}$$

[그림 2-10] Brier skill score 식

제3절 기계학습 모델의 학습과 검증

일반적인 통계분석과 기계학습을 이용한 예측모델 구축과의 가장 큰 차이점 중 하나는 그 목적에 있다. 일반적으로 통계분석의 목적은 전체 모집단에서 변수들의 관계를 표본집단을 분석하여 추정하는 데 있다. 예를 들어 흡연과 폐암 발생과의 관계를 전체 인구에서 확인할 수 없기에, 표본으로 얻은 코호트에서 이 관계를 확인하여 일반적인 관계를 추정하는 것이다. 반면, 기계학습은 미래에 일어날 사건을 예측하거나 새로운 것을 만들어 내기 위한 모델을 현재 가지고 있는 데이터로 만들고 이후 새로운 상황에 적용하고자 하는 것이다. 따라서 현재 가지고 있는 데이터셋에서의 모델의 성능보다, 앞으로 새로 접하게 될 데이터셋에 적용하였을 때 좋은 성능을 보일지 가늠하는 것이 훨씬 더 중요하다. 그러나 실제 미래에 일어날 일을 예측하여 데이터셋을 만들 수는 없으므로, 현재

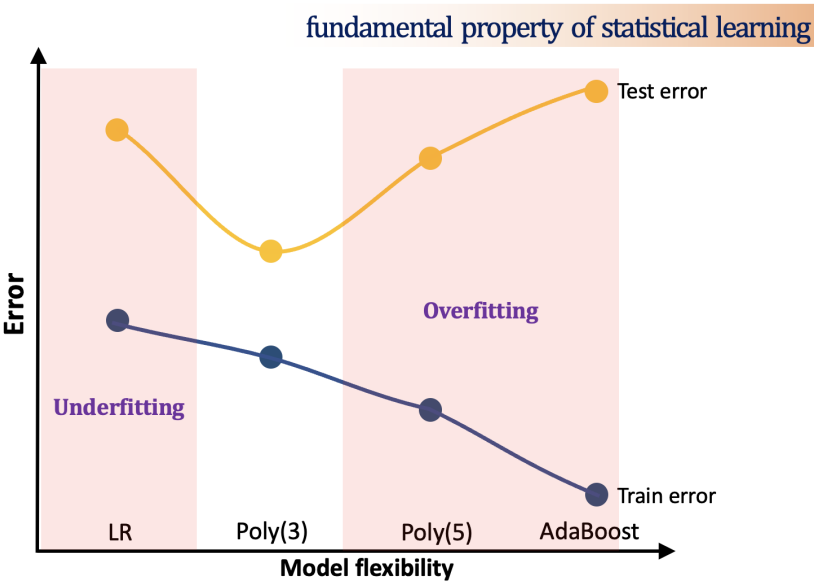
데이터셋 중 학습에 사용하지 않은 데이터를 학습된 모델을 최종 검증하여 확인하는 방식을 취한다(그림 2-1).



[그림 2-11] 일반적인 기계학습 예측모델의 학습 및 검증 과정을 나타낸 모식도

기계학습 알고리즘은 가지고 있는 학습용 데이터셋의 질과 양, 그리고 해결하고자 하는 과제의 난이도 등에 따라 그 복잡도가 달라져야 한다. 예를 들어, 어떤 사람의 소득수준이 어느 정도 예측하려고 하는 것에 비해, 그 사람이 모니터 앞에서 움직이는 영상을 분석하여 특정 질병 유무를 판단하려고 하는 것이 훨씬 더 어려운 과제이며, 더 많은 데이터와 더 복잡한 알고리즘이 필요하다. 최근 딥러닝의 붐이 일어난 것도, 이전에는 가능하지 않았던 복잡한 알고리즘이 개발되고 그 알고리즘을 학습시키고 적용할 수 있는 컴퓨터기술의 발전이 동반되었기 때문이다.

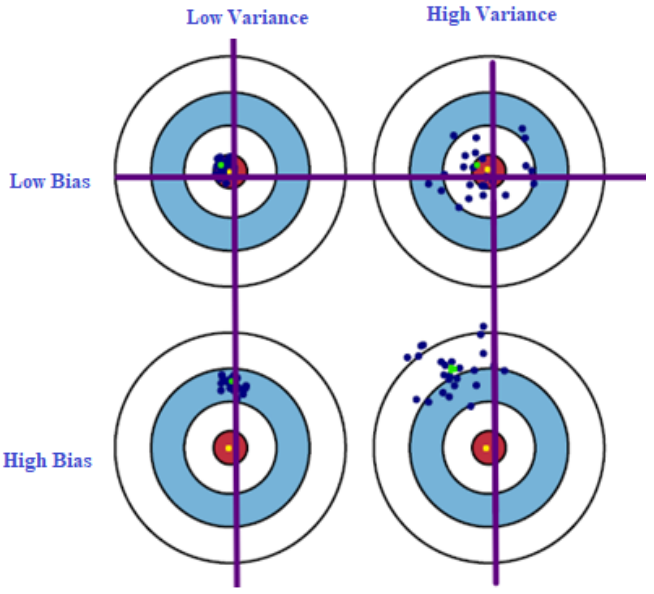
하지만 반대로, 주어진 과제에 비해 너무 복잡한 알고리즘을 이용하거나 충분한 데이터가 확보되지 않은 상태에서 알고리즘을 학습시키는 경우에는, 진짜 의미있는 신호 혹은 패턴을 데이터에서 찾아내지 못하고 특정 데이터셋에만 존재하는 잡음과 같은 신호를 의미있는 것으로 받아들인 채로 학습이 될 수 있다. 이렇게 되면 그 잡음이 있는 데이터셋에서는 성능이 괜찮지만 다른 데이터셋에 적용되었을 때는 낮은 성능을 보이게 된다. 이러한 현상을 모델이 특정 데이터셋에만 과하게 적합되었다 하여 과적합(overfitting)이라고 한다(그림 2-12).



모델의 복잡성 혹은 유연성이 증가할수록 과적합이 일어남을 보여줌. LR, Poly(3), Poly(5), AdaBoost로 갈수록 더 복잡하고 유연한 모델이며, 이에 따라 학습할 때의 오류는 계속 줄어들지만 검증할 때의 오류는 감소하다가 다시 증가하는 것을 볼 수 있다. 이 경우에는 Poly(3)가 최적의 모델이다.

[그림 2-12] 모델 유연성과 예측오류와 관계

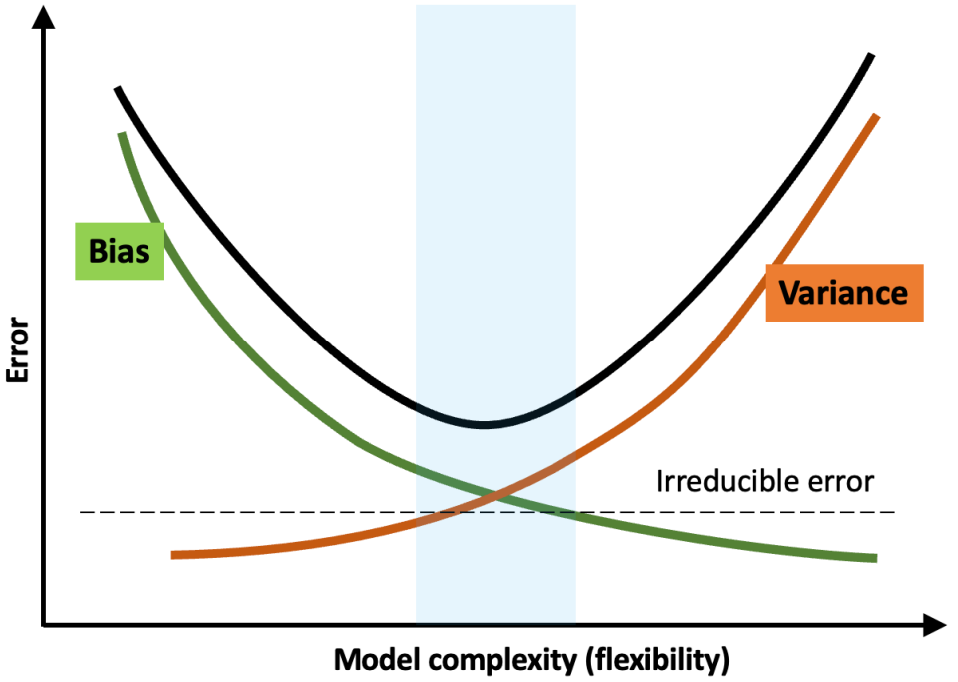
이와 관련하여 기계학습에서 중요한 개념이 bias와 variance이다. 둘 다 예측 모델의 오류를 나타내지만 내포하는 의미는 크게 다르다. [그림 2-13]은 기계학습 모델의 예측 결과를 bias와 variance 관점에서 해석하는 그림으로, 붉은 색 영역은 실제 값, 즉 정답을 의미하고 파란 점은 모델이 예측한 값을 의미한다. 여기서 bias는 정답과 예측 값의 차이를 의미하고, variance는 예측 값들의 흩어진 정도를 의미한다. 이 예시에서 직관적으로 둘 다 작은 (a) 모델이 가장 좋은 모델인 것을 알 수 있으며, (b) 모델은 예측 값들을 평균한 값은 정답 값과 비슷하지만(즉 bias가 작음), 예측 값들의 variance가 커서 오류가 큰 모델이다. 반면, (c) 모델은 bias가 크고 variance가 작은 모델이며, (d)는 둘 다 큰 모델이다.



[그림 2-13] Bias와 Variance

위 그림을 이제 학습과 검증의 관점에서 살펴본다면, 학습 시 사용한 데이터셋에서는 (a)와 같이 오류가 적었는데 검증 시에는 (b), (c), 혹은 (d)의 결과가 나올 수 있다. 세 가지 경우 모두 오류가 큰 것은 동일하지만 서로 원인은 다르다. Variance가 큰 (b) 모델은 모델이 학습용 데이터셋에 앞서 설명한 과적합이 된 것이 원인이다. Bias가 큰 (c) 모델은 반대로 모델이 해결하려고 하는 과제에 비해 충분히 복잡하지 않은 것이 원인이며, (d)는 둘 다의 경우로 생각할 수 있다.

중요한 점은 이 bias와 variance는 trade off 관계, 즉 하나를 높이면 필연적으로 나머지 하나는 낮아진다는 점이다(그림 2-14). 학습 시 높은 성능을 얻기 위해 모델의 복잡도를 높이는 것은 학습용 데이터셋에서의 bias를 줄이는 것이다. 그러나 이렇게 하면 여러 다른 데이터셋에 적용하였을 때, bias가 높지만 다른 데이터셋에도 비슷하게 적용될 수 있는 모델과 비교하여 상대적으로 데이터셋에 따라 결과가 계속 다르게 나오게 된다. 즉 variance가 높아지는 것을 의미한다. 따라서 한 관점에서는 기계학습 모델을 만든다는 것은 이 bias-variance tradeoff 하에서 최적의 모델을 만들기 위한 노력을 하는 것으로 볼 수도 있다.



[그림 2-14] Bias-Variance tradeoff

제3장

연구 자료 및 분석 방법

제3장
연구 자료 및 분석 방법

제1절 KCD 질병분류 및 조작적 정의

본 연구에서 이용한 건강검진표본코호트(NHIS-HEALS)는 한국표준질병사인분류(Korean Classification of Diseases, KCD) 6판에 근거하여 질병코드가 기록되어 있다. 한국표준질병사인분류는 국제표준 질병사인분류(International Classification of Diseases, ICD)를 근간으로 하고 있으며, 우리나라는 ICD 10판을 번역하여 1995년에 KCD-3을 고시한 후 계속하여 개정해 나가고 있다.[15,16] KCD-6의 기본구조와 원칙을 살펴보면 대분류, 중분류, 소분류, 세분류, 세세분류로 나눠져 있으며 첫 자리는 알파벳이고 둘째, 셋째, 넷째자리는 숫자로 되어 있다. 기본분류로는 3단위로 되어 있으며 최고 10개까지 4단위 분류로 세분할 수 있으나 표기 시는 소수점을 찍은 후 사용 가능하다.

본 연구에서는 국민건강보험공단의 청구자료 및 국가검진 결과에서 추출한 정보를 바탕으로 입력 변수들을 구성하였다. 해당 정보에는 나이, 성별, 소득분위 등의 인구사회학적인 정보, 의료이용 내역을 바탕으로 하여 재구성한 기저질환 및 복약 정보 등이 있다. 기저질환 유무 판단은 기본적으로 청구 내역에 포함된 KCD-6에 의한 진단명에 근거하지만, 입력된 진단명만으로 기저질환 유무를 판단해서는 정확하지 않다. 왜냐하면 의료 제공자가 청구할 때 입력한 진단명은 실제 환자의 진단명과 차이가 있을 수 있기 때문이다. 따라서 청구자료를 이용한 연구를 할 때 기저질환 유무는 조작적 정의를 이용하여야 한다.

따라서 본 연구에서는 이전 연구들에서 사용한 조작적 정의를 이용하여 기저질환 유무를 판단하였다.[17] 예를 들어, 간세포암은 간세포암에 해당하는 KCD-6 진단코드로 실제 입원하여 치료 받은 경우에만, 고혈압의 경우에는 KCD-6에 의한 고혈압 진단명과 함께 실제 고혈압 처방이 이루어지고 외래 방문이 2회 이상 이루어진 경우에만 실제 진단을 받은 것으로 정의하였다. 본 연구에서 사용한 모든 조작적 정의는 부록에 있는 <표 1>에서 찾아볼 수 있다.

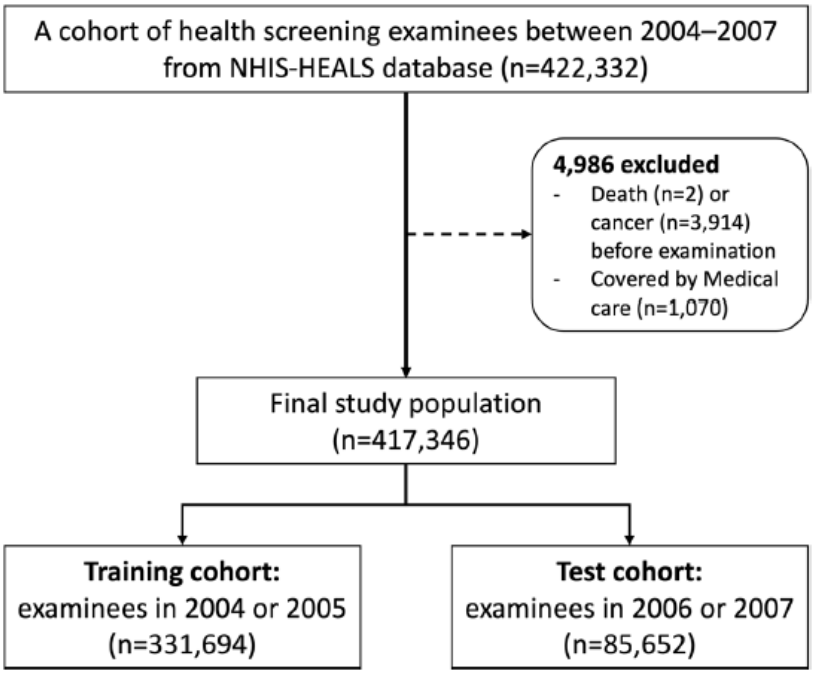
국민건강보험공단에서 제공하는 건강검진표본코호트에서는 일부 민감할 수 있는 진단명은 가려져 있다. 예를 들어, 조현병은 별개의 코드로 구분되어 알 수 없고 약물의존증 및 망상증과 결합되어 "F_"로 가려져 있다(부록표 1). 또한, 본 연구에서는 남성 혹은 여성에만 주로 해당하는 질환들인 전립선암과 유방암은 포함시키지 않았다.

제2절 연구 대상자

국민건강보험공단에서 제공하는 건강검진표본코호트(NHIS-HEALS)는 2002년과 2003년에 대한민국에서 국가건강검진을 수진한 40-80세 성인 남녀의 10%를 무작위로 추출한 51,4795명의 표본코호트이다. 포함된 사람들의 2002년부터 2015년 사이의 청구내역과 건강검진결과가 포함되어 있다.

표본코호트에 포함된 수진자 총 51,4795명 중 41,7346명이 연구에 포함되었다. 2002년과 2003년을 씻김시기(washout period)로 두고 이 시기를 포함하여 검진 전 시기에 일어난 일을 바탕으로 각 연구대상자의 기저질환 등을 파악하였다. 이 씻김시기에 사망하거나 암을 진단받은 사람들 3,916명은 연구에서 제외하였다. 또한, 의료보호 대상자 1,070명도 그 수가 적고 일반 의료보험가입자와 의료이용 행태가 크게 달라 추가로 연구에서 제외하였다(그림 3-1).

알고리즘을 학습시키고 검증하기 위해, 2004년 혹은 2005년에 검진을 받은 33,4966명은 기계학습 알고리즘을 학습시키는 데 사용되는 "학습군(training set)"에, 2004년 혹은 2005년에는 검진받지 않았으나 2006년 혹은 2007년에 검진을 받은 8,7416명은 학습한 알고리즘을 검증하기 위한 "테스트군(test set)"에 속하도록 대상자를 두 군으로 나누었다. 이렇게 나누어진 두 군의 대상자 수 비는 대략 8대 2로, 대부분의 기계학습 모델 수립 시 채택하는 비인 7:3 혹은 8:2에 합당한 결과임을 확인하였다.[18]



[그림 3-1] 연구 대상자 선정과정

제3절 입력변수와 결과변수

1. 입력변수

기계학습 모델을 수립할 때, 무작정 많은 변수들을 이용하는 것은 쓸모없는 정보가 유입되어 알고리즘이 잡음을 중요한 신호로 판단하여 일반화가 잘 안되는 모델로 이어질 수 있다.[19] 이렇게 되면 학습할 때 이용한 데이터로는 좋은 성능을 보이지만 처음 접하는, 즉 해당 잡음과 다른 신호를 가진 데이터에 적용되었을 때는 형편없는 성능을 보일 수 있다. 이를 과적합이 일어났다고 한다. 따라서 사전에 의미없는 데이터를 제거하고 학습을 시키는 것은 좋은 모델을 만들고 학습시키는 데 도움이 된다.

이러한 노력 중 하나가 입력변수를 정할 때 어느 정도 사전작업을 거치는 것이다. 이를 위해 본 연구에서는 먼저 콕스 비례-위험 회귀분석을 통해 의미 있는 변수들만 선택하였다.

먼저, 각 변수들의 공선성을 분석하여 분산팽창계수(variation inflation factor, VIF)가 2.5 이상인 변수들을 골라내었다. 본 연구에서는 수축기/이완기 혈압과 aspartate

transaminase (AST)/alanine transaminase (ALT) 수치가 서로 강한 상관성을 나타내었다(부록표 2). 따라서 혈압은 수축기 혈압과 이완기 혈압의 평균을 내어 사용하였고 ALT가 더 간질환에 특이적이므로 AST는 버리고 ALT만 변수로 사용하였다.

다음으로, 단변수 콕스 비례-위험 회귀분석에서 p 값이 0.05 미만으로 간세포암 발병과 의미있는 관계를 보였던 변수들만으로 다변수 회귀분석에 독립적으로 유의한 관계를 보인 변수들만 의미있는 예측인자로 선택하였다. 단순히 우연히 유의한 인자로 선택되는 것을 최소화하기 위해, 학습군 데이터를 부트스트랩(bootstrap) 1000번으로 1000개의 서로 다른 데이터셋에서 이 과정을 반복하였고, 여기서 최소 850번 이상 유의한 관계를 보였던 변수들만 최종 예측인자 변수로 채택하였다.

2. 결과변수

본 연구에서는 크게 두 가지 관점에서 간세포암 발병위험을 예측하였다. 첫째는 9년 이상 관찰하는 동안 최종적으로 간세포암이 발병할지 아니면 발병하지 않을지를 이분법적으로 예측하는, "분류(classification)"의 관점이었다. 둘째는 만약 검진 수진자가 간세포암에 걸린다면 검진 후 얼마 후에 발병할지 그 시기를 예측하는 "발생까지 시간을 예측(time-to-event prediction)"하는 관점이었다.

건강검진표본코호트(NHIS-HEALS)는 2015년까지의 추적관찰 정보를 제공하므로, 발생까지 시간을 예측하는 데 있어서 2015년 12월 31일을 마지막 관찰일로 하였고 이전에 간세포암이 발병하지 않은 경우에는 우측절단(right censored) 처리하였다. 건강검진표본코호트에 포함된 사망원인과 사망일은 대한민국 통계청에서 제공하는 정보가 결합되어 있는 것이다.

제4절 기계학습모델의 학습

1. 통계분석 및 기계학습

선택된 최종 입력변수를 이용하여, 다변수 콕스 비례-위험 생존분석에서 위험률(hazard ratio, HR)를 계산하였다. 간세포암 외 다른 암종의 경우도 간세포암과 비슷한 원인인자를 공유할 수 있으므로, 추가로 간세포암외 다른 암종을 경쟁자로 두고 보정하여 분석하는 경쟁위험(competing risk) 분석을 추가로 진행하였다.[20]

연속형 변수의 비교는 Mann-Whitney 검정 혹은 t 검정을, 범주형 변수의 비교는 카이 검정을 이용하였다. 연속형 변수는 다른 언급이 없으면 평균과 표준편차로 대푯값

을 나타내었다. 양측 확률값으로 0.05 미만인 경우 유의한 관계 혹은 차이가 있다고 간주하였다. 모든 통계분석과 기계학습은 R 3.3.3을 이용하여 수행하였다. 사용한 패키지는 다음과 같다: survival (v2.41-3), cmprsk (v2.2-7), randomForestSRC (v2.5.1), caret (v6.0-78), survminer (0.4.2), 'xgboost (0.6.4.1).

2. 학습군에서 기계학습모델의 학습

생존 랜덤포레스트 알고리즘은 "분류"와 "발생까지 시간 예측"에 모두 사용되었다.[21] 모델 학습 시 간세포암 외 다른 암종을 경쟁위험으로 간주하였다. 학습군에서 10-분절 교차검증(10-fold cross validation)을 통해 파라미터의 최적화를 진행하였다. 이 과정을 통해 결정된 최적의 파라미터는 ntree = 120, mtry = 1, nodesize = 6이었다.

XGBoost 알고리즘은 간세포암 발병 유무를 결정하는 "분류" 과제에 이용되었다. 생존 랜덤포레스트 알고리즘의 경우와 마찬가지로 학습군에서 10-분절 교차검증(10-fold cross validation)을 통해 파라미터의 최적화를 진행하였다. 이 과정을 통해 결정된 최적의 파라미터는 max.depth = 5, eta = 0.1, min_child_weight = 1, gamma = 0, lambda = 0, nrounds = 108이었다.

최적의 파라미터들을 찾은 후에는 이 파라미터들를 설정한 모델을 다시 전체 학습군에서 최종학습시키고, 같은 데이터셋에서 학습된 콕스 비례-위험 회귀모형과 성능을 비교하였다. 더 나아가, 간세포암 발병 예측에 있어서 생존 랜덤포레스트와 XGBoost 알고리즘의 결합한 앙상블 모델을 수립하여 각 알고리즘을 독자적으로 사용하였을 때와 비교해 성능 향상이 있을지 확인하여 보았다.

3. 평가지표

간세포암 발병시기까지의 기간을 예측하는데 있어서 사용된 평가지표는 크게 3가지로, 분별력(discriminative ability), calibration, 그리고 전반적인 정확도(overall accuracy)이다. 각각 c-index, calibration 도표, 그리고 Brier skill score를 이용하여 평가하였다. 각 지표의 설명은 제2장 이론적 고찰에서 다루었다. 발병 유무를 판단하는데 있어서는 진단적 정확도, 민감도, 특이도, AUC를 평가지표로 이용하였다.

제5절 기계학습모델의 검증

최종 예측모델을 테스트군에서 검증하였다. 평가지표는 모델 학습 시와 마찬가지로 발병시기까지의 기간을 예측하는데 있어서는 c-index, calibration 도표, Brier skill score를, 발병 유무를 판단하는 데 있어서는 정확도, 민감도, 특이도, AUC를 평가지표로 이용하였다.

예측된 간세포암 발병 위험 가능성에 따라 수진자를 저위험군 (<5%), 중위험군 (5-20%), 고위험군(>20%)의 세 군으로 나누고 Kaplan-Meier 생존곡선을 그리고 각 군의 생존 가능성에 유의한 차이가 있는지 로그-랭크 검정을 통해 확인하였다. 위험군을 세 군으로 나누는 것은 각 군의 대상자 수가 비슷해지도록 임의로 기준을 정한 것이다.

마지막으로, 전술한 방법과 같은 방법으로 당뇨병이 있는 환자군, 알콜성 지방간이 있는 환자군, 그리고 비알콜성 지방간이 있는 환자군에서 분석을 진행하여, 해당 질환이 있는 환자군에서의 결과 역시 확인하였다.

제4장

분석 결과

제4장
분석 결과

제1절 연구 대상자

최종 연구대상자는 41,7346명이었고 79.5%인 33,1694명이 학습군, 나머지 20.5%인 8,5652명이 테스트군에 속하였다. 건강검진 수진일 기준으로 전체 대상자의 평균 나이는 55세였고 그 범위는 42세에서 82세까지였다. 남성 대 여성의 비는 약 5.5:4.5로 남성이 좀 더 많았다. 거의 대부분의 변수들이 학습군과 테스트군 사이에 유의한 차이가 있었고 이는 두 군의 성질이 다소 다르다는 것을 보여준다(표 4-1과 부록표 3).

<표 4-1> 학습군과 테스트군에 속하는 연구 대상자들의 특성

변수		학습군	테스트군	p-value	전체
인구사회학적 특성					
나이		54.3 (9.28)	57.7 (9.6)	<0.001	55.0 (9.45)
성별	여성	42.2% (140035/331694)	55.5% (47546/85652)	<0.001	44.9% (187581/417346)
	남성	57.8% (191659/331694)	44.5% (38106/85652)		55.1% (229765/417346)
소득수준	<30%	28.5% (94614/331694)	30% (25732/85652)	<0.001	28.8% (120346/417346)
	30-80%	34.9% (115713/331694)	40.5% (34656/85652)		36% (150369/417346)
	>80%	36.6% (121367/331694)	29.5% (25264/85652)		35.1% (146631/417346)
신체계측					
체질량 지수	정상	66.2% (219447/331529)	64.2% (54942/85617)	0.213	65.8% (274419/417217)
	과체중	31.3% (103922/331529)	32.6% (27933/85617)		31.6% (131855/417217)
	비만	2.5% (8130/331529)	3.2% (2742/85617)		2.6% (10943/417217)

변수		학습군	테스트군	p-value	전체
혈압	수축기	126.59 (17.19)	126.54 (17.17)	0.33	126.58 (17.18)
	이완기	79.18 (11.15)	78.37 (10.82)	<0.001	79.02 (11.09)
혈액검사					
AST (IU/L)		26.55 (15.92)	26.47 (17.37)	0.25	26.53 (16.23)
ALT (IU/L)		25.52 (19.45)	24.9 (20.63)	<0.001	25.39 (19.7)
GGT (IU/L)		37.7 (51.85)	36.58 (53.85)	<0.001	37.47 (52.27)
총콜레스테롤 (mg/dL)		198.32 (36.84)	199.45 (37.84)	<0.001	198.55 (37.05)
금식혈당 (mg/dL)		97.87 (28.9)	99.53 (28.12)	<0.001	98.21 (28.75)
혈색소 (g/dL)		13.94 (1.49)	13.65 (1.51)	<0.001	13.89 (1.5)
생활습관					
흡연(팩-년)		5.96 (11.53)	4.6 (11.11)	<0.001	5.68 (11.46)
음주(ml/주)		8.7 (18.7)	7.7 (19.1)	<0.001	8.5 (18.8)
운동	거의 안함	50.1% (162668/324506)	56.4% (46716/82888)	<0.001	51.4% (209384/407394)
	주1-2회	26.9% (87443/324506)	22.2% (18437/82888)		26% (105880/407394)
	주3-4회	12.1% (39341/324506)	10.4% (8647/82888)		11.8% (47988/407394)
	주5-6회	3.2% (10384/324506)	3.1% (2587/82888)		3.2% (12971/407394)
	거의 매일	7.6% (24670/324506)	7.8% (6501/82888)		7.7% (31171/407394)
가족력					
간질환		2.81% (8563/305086)	2.64% (2040/77356)	0.01	2.77% (10603/382442)
고혈압		9.16% (28072/306539)	9.69% (7548/77870)	<0.001	9.27% (35620/384409)
뇌졸중		5.47% (16738/305817)	5.55% (4310/77600)	0.38	5.49% (21048/383417)
심장질환		2.39% (7278/305106)	2.57% (1992/77365)	<0.001	2.42% (9270/382471)
당뇨병		6.45% (19729/306048)	6.8% (5280/77653)	<0.001	6.52% (25009/383701)
암		13.14% (40389/307443)	13.45% (10498/78065)	0.02	13.2% (50887/385508)

변수	학습군	테스트군	p-value	전체
기저질환				
당뇨병	6.07% (20139/331694)	8.83% (7565/85652)	<0.001	6.64% (27704/417346)
고지혈증	5.95% (19725/331694)	10.45% (8950/85652)	<0.001	6.87% (28675/417346)
만성간염바이러스	2.57% (8530/331694)	3.71% (3176/85652)	<0.001	2.8% (11706/417346)
HIV 감염증	6.73% (22314/331694)	10.6% (9079/85652)	<0.001	7.52% (31393/417346)
조현병 등 정신질환	15.59% (51714/331694)	25.57% (21905/85652)	<0.001	17.64% (73619/417346)
만성간질환	5.4% (17927/331694)	8.21% (7033/85652)	<0.001	5.98% (24960/417346)
알콜성 지방간	1.95% (6480/331694)	2.75% (2353/85652)	<0.001	2.12% (8833/417346)
비알콜성 지방간	3.01% (9996/331694)	5.07% (4345/85652)	<0.001	3.44% (14341/417346)

이 표에는 기저질환 모두가 포함되어 있지 않음. 모든 기저질환에 대한 결과는 부록의 표 3을 참고.
AST=aspartate aminotransferase, ALT=alanine aminotransferase, GGT=gamma-glutamyl transferase. HIV=human immunodeficiency virus.

추적관찰기간의 중간값은 학습군에서 11.1년(최대 12.0년까지)이었고 테스트군에서 9.1년(최대 10.0년까지)였다. 학습군 33,1694명 중 0.5%인 1799명과 테스트군 8,5652명 중 0.5%인 390명이 관찰기간 중 간세포암이 발병하였다. 간세포암 외 다른 암종은 학습군 33,1694명 중 8.4%인 27856명, 테스트군 8,5652명 중 7.9%인 6732명에서 관찰기간 중 발생하였다.

제2절 예측인자

1,000번의 부트스트랩 표본에서 85% 이상 간세포암 발병과 유의한 관계를 보였던 변수들은 다음과 같다(부록표 4). 인구사회학적 특성 중에는 나이, 성별, 비만, 소득수준, 만성간질환의 가족력이었다. 검진결과에서는 ALT, gamma-glutamyl transferase (GGT), 혈중 총콜레스테롤 수치였다. 기저질환 중에는 만성간염바이러스 감염력, 인간면역결핍바이러스(human immunodeficiency virus, HIV) 감염력, 당뇨병, 고지혈증, 그리고 조현병 등 정신질환이 선택되었다.

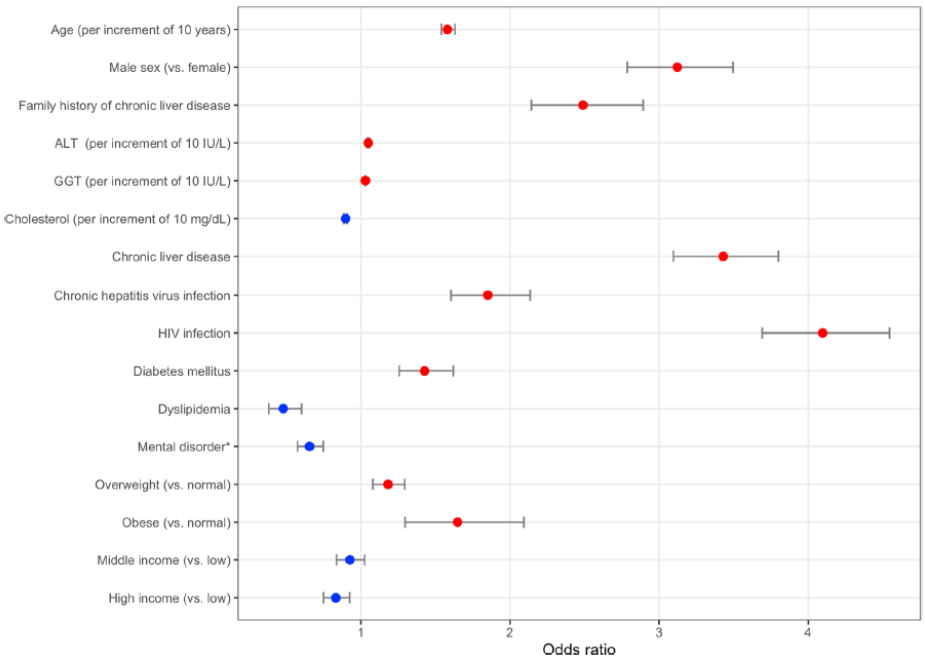
다변수 콕스 비례-위험 회귀분석에서, 더 나이가 많을수록(hazard ratio [HR], 1.581 / 10살), 여성보다 남성에서(HR, 3.122), 만성간질환의 가족력이 있을수록(HR, 2.490), 비만인 경우(HR, 1.648), ALT 수치가 높은 경우(HR, 1.049 / 10 IU/L), GGT 수치가 높은 경우(HR, 1.030 / 10 IU/L), 만성간질환이 있는 경우(HR, 3.430), 만성간염바이러스감염력이 있는 경우(HR, 1.851), HIV 감염력 있는 경우(HR, 4.097), 당뇨병이 있는 경우(HR, 1.427)에 간세포암 발병위험이 더 높았다. 반면, 혈충 총 콜레스테롤이 높은 경우(HR, 0.897 / 10 mg/dL), 고지혈증이 있는 경우(HR, 0.479), 조현병 등 정신질환이 있는 경우(HR, 0.655), 소득수준이 높은 경우(HR, 0.832)는 간세포암 발병위험이 더 낮았다(모든 경우 p <0.001). 간세포암 외 다른 암종을 경쟁인자로 두고 분석하였을 때도 비슷한 결과를 얻었다(표 4-2와 그림 4-1). 즉, 다시 말해, 다른 암 발생이 간세포암 발병에 미치는 영향을 보정한 뒤에도 여전히 해당 인자들이 간세포암 발병 위험에 독립적으로 유의한 인자들임을 확인하였다.

<표 4-2> 간세포암 외 다른 암종을 경쟁위험으로 간주하지 않았을 때와 간주 하였을 때의 다변수 Cox 비례-위험 회귀분석 결과: 간세포암 발병위험과 유의한 독립적으로 유의한 관련성을 보인 변수들의 위험률 및 95% 신뢰구간

변수	Hazard ratio	95% 신뢰구간	p-value
다른 암종을 경쟁인자로 고려하지 않을 때			
나이	1.581	1.540-1.629	<0.001
남성(vs. 여성)	3.122	2.786-3.496	<0.001
ALT	2.490	2.143-2.893	<0.001
GGT	1.049	1.044-1.054	<0.001
총콜레스테롤	1.030	1.027-1.032	<0.001
만성간질환	0.897	0.886-0.908	<0.001
만성간염바이러스 감염력	3.430	3.096-3.800	<0.001
HIV 감염력	1.851	1.605-2.135	<0.001
당뇨병	4.097	3.691-4.546	<0.001
고지혈증	1.427	1.257-1.619	<0.001
조현병 등 정신질환	0.479	0.382-0.601	<0.001
과체중(vs. 정상체질량지수)	0.655	0.575-0.747	<0.001
비만(vs. 정상체질량지수)	1.182	1.080-1.294	<0.001
중간 소득수준(vs. 낮은)	1.648	1.297-2.094	<0.001
높은 소득수준(vs. 낮은)	0.926	0.836-1.025	0.138

변수	Hazard ratio	95% 신뢰구간	p-value
다른 암종을 경쟁인자로 고려할 때			
나이	1.542	1.493-1.581	<0.001
남성(vs. 여성)	3.040	2.710-3.401	<0.001
ALT	2.487	2.099-2.948	<0.001
GGT	1.048	1.036-1.061	<0.001
총콜레스테롤	1.029	1.026-1.032	<0.001
만성간질환	0.898	0.887-0.910	<0.001
만성간염바이러스 감염력	3.470	3.116-3.863	<0.001
HIV 감염력	1.862	1.567-2.213	<0.001
당뇨병	4.057	3.640-4.522	<0.001
고지혈증	1.430	1.255-1.629	<0.001
조현병 등 정신질환	0.458	0.359-0.585	<0.001
과체중(vs. 정상체질량지수)	0.645	0.563-0.738	<0.001
비만(vs. 정상체질량지수)	1.190	1.084-1.308	<0.001
중간 소득수준(vs. 낮은)	1.662	1.304-2.118	<0.001
높은 소득수준(vs. 낮은)	0.925	0.831-1.030	0.150

ALT = alanine aminotransferase, GGT = gamma-glutamyl transferase, HIV = human immunodeficiency



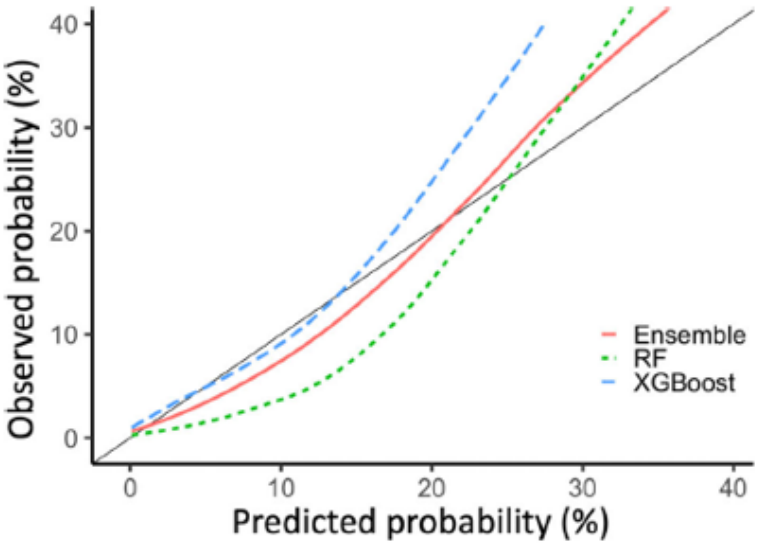
[그림 4-1] 각 변수의 교차비(odds ratio)를 나타낸 Forest 도표

제3절 기계학습

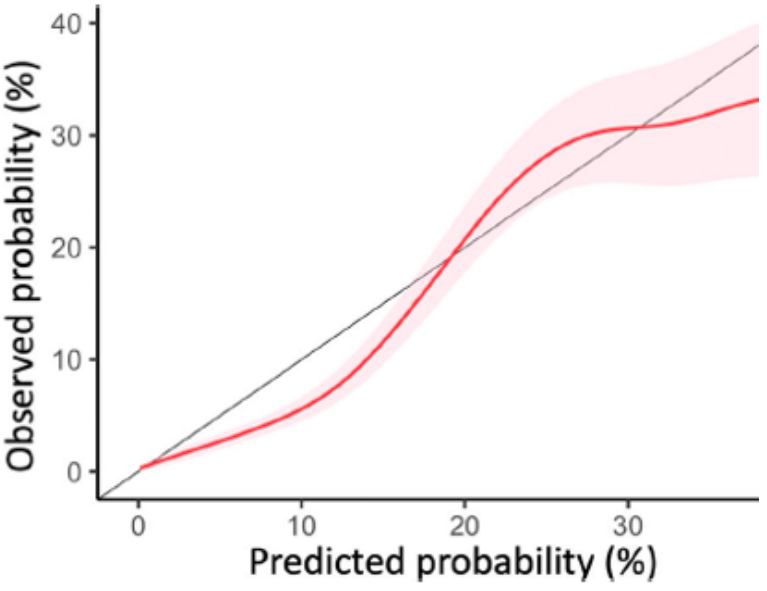
1. 관찰기간 내 간세포암 발병 여부 예측

학습군에서, 간세포암 발병 위험을 예측하는데 있어 XGBoost가 생존랜덤포레스트보다 더 우수한 성능을 보였다. 간세포암이 관찰기간 내 발병할지 아니면 하지 않을지 이분법적으로 판단하는데 있어서, AUC (+/- 표준편차)는 XGBoost와 생존랜덤포레스트가 교차검증에서 각각 0.882 (+/-0.013)과 0.871 (+/-0.019)이었다(표 4-3). Calibration 도표에서 두 알고리즘은 비슷하게 좋은 성능을 나타내었다(그림 4-2). Brier skill score는 각각 0.109와 0.062로 이는 기저평가 기준인 모든 대상자에서 간세포암이 발병하지 않을 것이라 예측한 경우와 비교하여 XGBoost와 생존랜덤포레스트 모델을 사용하였을 때 각각 10.9%와 6.2%의 Brier score 향상이 예상된다는 의미이다. 두 알고리즘의 앙상블 모형이 가장 좋은 성능을 보였는데, AUC, Brier skill score는 각각 0.892 (+/-0.011)과 0.112였다. 따라서 앙상블 모델이 우리의 최종모델로 선택되었다.

최종적으로 학습된 예측모델을 테스트군에서 최종 검증하였다. 20% 이하 확률에서 다소 위험을 과소평가하는 경향을 보였으나 전반적으로 좋은 calibration을 보였다(그림 4-3). AUC는 0.873이었고 95% 신뢰구간은 0.860-0.885였다. Brier skill score는 0.078로 기저평가 기준에 비해 약 7.8% 정도 Brier score의 향상이 있었다. 발병확률 1%을 기준으로 그 이상인 경우 발병 위험이 있다고 보았을 때의 진단적 민감도, 특이도, 정확도는 각각 71.8% (95% 신뢰구간, 71.4-72.2), 88.4% (95% 신뢰구간, 88.1-88.7), 88.4% (95% 신뢰구간, 88.2-88.6)이었다.



[그림 4-2] 학습군에서의 calibration 도표. 초록색 점선은 생존랜덤포레스트 모델(RF), 파란색 점선은 XGBoost, 빨간색 곡선은 두 모형을 앙상블한 모델을 나타낸다.



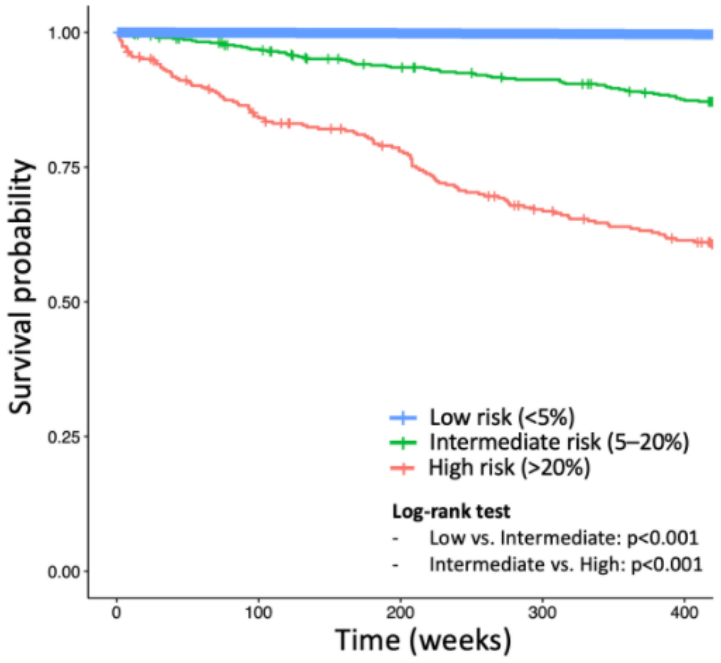
[그림 4-3] 테스트군에서의 최종 앙상블 모델의 calibration 도표. 왼쪽

<표 4-3> 기계학습 모델의 성능평가 결과

모델	평가지표	교차검증결과	최종검증결과
관찰기간 내 간세포암이 발병할 확률을 예측			
생존랜덤포레스트	AUC	0.871 (±0.019)	
	BSS	0.062	
XGBoost	AUC	0.882 (±0.013)	
	BSS	0.109	
앙상블	AUC	0.892 (±0.011)	0.873 (0.860-0.885)
	BSS	0.112	0.078
간세포암 발병까지 시간을 예측			
Cox 비례-위험모형	C-index	0.843 (±0.006)	0.828 (0.819-0.838)
생존랜덤포레스트	C-index	0.881 (±0.010)	0.857 (0.850-0.864)

괄호 내 값은 교차검증결과의 경우는 표준편차, 최종검증결과의 경우는 95% 신뢰구간임.
AUC = area under receiver operating characteristics curve, BSS = Brier skill score, C-index = concordance index.

Kaplan-Meier 곡선에서 대상자를 세 위험군, 즉 저위험군(<5%), 중위험군(5-20%), 고위험군(>20%)으로 나누었을 때, 차례로 유의하게 발병위험이 증가하는 결과를 보였다(모든 비교에서 p <0.001; 그림 4-4).

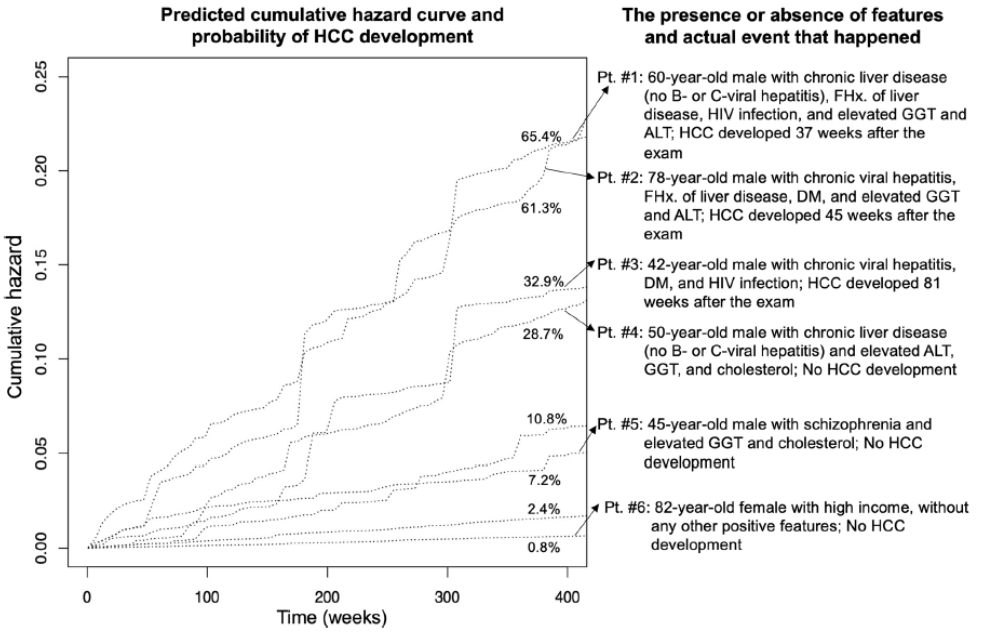


[그림 4-4] 간세포암 발병위험에 따라 세 위험군으로 나누었을 때의 생존곡선
(저위험군, <5%; 중위험군, 5-20%; 고위험군, >20%)

2. 간세포암 발병까지 시간 예측

간세포암 발병까지 걸린 시간의 중간값은 학습군에서 약 294주 혹은 5.6년, 테스트군에서 약 235주 혹은 4.5년이었다. 발병까지 걸린 시간을 예측하는 데 있어서 생존랜덤포레스트 모델이 콕스 비례-위험 모델보다 우수한 성능을 보였다(표 4-3). 테스트군에서, 콕스 모델은 AUC가 0.828 (95% 신뢰구간 0.819-0.838)이었던 반면 생존랜덤포레스트 모델은 0.857 (95% 신뢰구간 0.850-0864)로 95% 신뢰구간이 겹치지 않아 유의한 차이가 있는 것으로 해석되었다.

이러한 간세포암 발병 위험은 기계학습 모델에 의해 개인별로 그 결과를 확인할 수 있다. 그림 4-5에 그 예를 나타내었다. 각 개인별로 대략 10년 내 간세포암이 발병할 전반적인 위험도(백분률로 나타냄)와, 시간이 지날수록 그 위험이 어떻게 증가하는지 알 수 있다. 발병위험이 높은 사람일수록 시간이 지날수록 위험이 증가하는 증가율이 높고, 낮은 사람일수록 시간이 지나도 위험이 크게 증가하지 않는다. 더 나아가, 어떠한 요인 때문에 위험증가가 예상되는지 알 수 있어, 가능한 경우 해당 위험인자를 줄이기 위한 노력을 하여 발병 예방에 도움이 될 수 있다. 예를 들어, 그림 4-5에서 첫 번째 수진자(Pt #1)의 경우 60세 남성이고 만성 간질환과 간질환 과거력이 있는 HIV 환자였으며 검진 결과 ALT와 GGT 상승이 관찰되어, 이 수진자의 10년 내 간세포암 발병 확률은 65.4%나 되었다. 이 분은 실제로 검진 후 37주 후 간세포암이 진단되었다. 반면, 여섯 번 째 수진자의 경우에는 82세의 고령이었으나 다른 위험인자가 없었고 검진 결과도 특이 사항이 없었다. 예측된 발병 위험은 1% 미만이었고, 관찰 기간 내 간세포암 발병은 없었다.



[그림 4-5] 6명의 예시 결과들

각 수진자의 검진 후 400주까지 간세포암이 발병할지 여부를 백분율로 그 위험을 보여주고, 시간이 지나면서 발병 위험이 어떻게 증가하는 보여준다.

3. 당뇨병 혹은 지방간이 있는 환자군에서의 결과

당뇨병, 알콜성 지방간, 그리고 비알콜성 지방간이 있는 환자군에서 각각 같은 분석을 하였을 때, 다소 감소했으나 전반적으로 비슷한 결과를 얻었다. 테스트군에서 검증한 결과, 최종 앙상블 모델의 AUC는 각각 0.851 (95% 신뢰구간, 0.794-0.862), 0.853 (95% 신뢰구간, 0.801-0.822), 그리고 0.849 (95% 신뢰구간, 0.837-0.861)이었다(표 4-4).

<표 4-4> 당뇨병 혹은 지방간이 있는 환자군에서 최종 앙승블 모델의 간세포암 발병위험 확률 예측 성능평가 결과

환자군	평가지표	교차검증결과	최종검증결과
당뇨병	AUC	0.873 (±0.006)	0.851 (0.794-0.863)
알콜성 지방간	AUC	0.882 (±0.006)	0.853 (0.801-0.822)
비알콜성 지방간	AUC	0.874 (±0.006)	0.849 (0.837-0.861)

괄호 내 값은 교차검증결과의 경우는 표준편차, 최종검증결과의 경우는 95% 신뢰구간임.
AUC = area under receiver operating characteristics curve.

제5장

결론

제5장 결론

본 연구를 통해 국가검진 후 약 10년 내 간세포암이 발병할 위험을 검진결과와 국민건강보험 청구자료에 근거하여 예측하는 기계학습 모델을 구축하고 검증하였다. 이 예측모델은 검증 결과 학습 시 예상했던 성능을 보임이 확인되었다. 더 나아가, 당뇨병 및 지방간이 있는 특정 환자군에서도 같은 결과를 얻었다.

이전에 출판된 모델들은 대부분 만성간질환이 이미 있어 간세포암 고위험군으로 분류되는 환자들을 대상으로 하는 모델들이었다. 이 예측모델들의 정리 및 설명은 이전 다른 문헌에 자세히 소개되어 있다.[11,12]

지금까지 출판된 예측모델들 중에는 현재까지 3개의 모델만이 만성간질환 유무에 관계없이 일반 인구를 대상으로 적용할 수 있다. Michikawa 등은 1,7654명의 건강검진 수진자의 일본인 코호트를 이용하여 나이, 성별, 알콜 혹은 커피 소비, 비만, 당뇨병, 만성 B형 혹은 C형 간염바이러스 감염이 일반적인 인구집단에서 간세포암 발병의 독립적 예측인자임을 보였고, 이를 입력변수로 하는 예측모형을 만들었다.[22] Wen 등은 42,8584명의 대만인 코호트를 이용하여 나이, 성별, 알콜 사용, ALT, AST, 알파태아혈청단백, 당뇨병, 만성 B형 혹은 C형 간염바이러스 감염이 유의한 위험인자임을 밝히고 간세포암 예측모형을 만들었다.[23] 최근 한국에서 본 연구와 마찬가지로 국민건강보험공단 국가검진표본코호트를 이용한 연구가 출판되었다.[24] 이 연구에서는 나이, 성별, 흡연, 당뇨병, 총콜레스테롤, ALT를 유의한 예측인자 및 입력변수로 하여 모델을 만들었다. 하지만 이 연구는 기존에 잘 알려진 위험인자인 간경변증과 만성간염바이러스감염이 없는 사람들만을 대상으로 하여 본 연구의 모델보다 그 대상범위가 작다. 본 연구의 모델의 경우에는 국가검진 후 수진자가 자신이 고위험군에 속하는지 아는 것과 상관없이 위험을 예측하여 보고하는 것이 목표였기 때문에, 기본적으로 전체 수진자를 대상으로 모델을 만들었다.

모든 이전 모델들은 Cox 비례-위험 회귀모형을 이용하여 예측모델을 만들었는데, 본 연구의 모델은 생존랜덤포레스트와 XGBoost의 기계학습 알고리즘을 이용하였고,

기계학습 알고리즘이 Cox 모델보다 우수할 수 있음을 보였다. 또한, 이전 모델들이 간염바이러스 감염력 등 잘 알려진 인자들만을 이용하여 모델을 만든 반면, 본 연구에서는 보험 청구자료와 검진 결과에서 얻을 수 있는 많은 변수들 중, 이전에 고려하지 않았으나 잠재적으로 위험인자가 될 수 있는 변수를 찾기 위해 많은 노력을 기울였다. 그 결과, 조현병 등 정신질환, 고소득 등, 기존 위험인자만으로 모델을 만들었다면 버려졌을 정보를 활용할 수 있었다.

그러나 우리는 잠재적 위험인자를 찾는 과정에 있어서, 실제로는 의미가 없으나 우연히 의미있는 것처럼 보이는 인자들을 선택하지 않도록 역시 많은 노력을 기울였다. 복잡한 기계학습 알고리즘은 주어진 자료 크기나 과제에 비해 너무 유연한 구조를 가짐으로써 사소하고 의미 없는 신호를 중요한 패턴이라 생각하는 함정에 빠질 수 있다. 이렇게 되면 학습할 때 이용한 특정 데이터셋에만 존재하는 잡음과 같은 신호를 의미 있는 것처럼 받아들이게 되어, 실전에서 다른 데이터를 접하였을 때는 엉뚱한 예측결과를 낼 수 있다(과적합, overfitting). 본 연구에서는 이러한 함정을 피하기 위하여 부트스트랩으로 1,000개의 서로 구성이 다른 데이터셋을 생성한 후 이를 대상으로 유의한 인자를 찾는 과정을 반복하였으며, 반복된 분석에서 85% 이상 유의한 관계를 보인 인자들만 최종 예측인자로 채택하였다. 본 연구에서도, 간세포암 발병과 그 관련성을 설명하기 어려운 치핵, 만성비염등이 생성된 데이터셋에서 자주 독립된 위험인자로 선택된 것이 이런 현상을 증명한다(부록표 4).

고령, 남성, 만성간질환, 과음, 당뇨병, 비만, HIV 감염은 이미 잘 알려진 감세포암의 위험인자이며,[25,26] 본 연구에서도 모두 유의한 위험인자임이 확인되었다. 한 가지 예외가 음주력인데, 본 연구에서는 최종 위험인자로 선택되지 않았다. 대부분 연구에서 음주는 간세포암 발병의 위험인자임이 확인되는 것에 반하는 결과이다. 그러나, 본 연구와 같은 코호트로 진행된 최근 연구에서도 본 연구와 마찬가지로 설문에 의해 확인되는 과음이 간세포암 발병과 유의한 관련성을 보이지 않았다.[24] 따라서 이는 국가검진 시 시행되는 설문에 의한 음주력 조사가 정확한 음주력을 반영하지 못할 수 있음을 시사한다고 해석해볼 수 있겠다.[27] 그 대신, ALT 수치는 이전에 알려진 대로 강력한 위험인자였는데, 혈액검사의 경우 객관적으로 확인할 수 있는 수치이므로 이것으로 음주를 포함한 다양한 원인에 의한 간손상을 간접적으로 파악할 수 있다고 볼 수 있다.

간암의 가족력은 간세포암의 위험인자로 알려져 있으나,[30,31] 본 연구에 사용된 설문에서는 암종을 구분하지 않고 모든 암을 포함하여 암 가족력을 조사하고 있다. 이것이

본 연구 분석 결과 암 가족력이 간세포암 발병 위험증가와 무관하였던 원인으로 사료된다.

당뇨병과 달리, 고지혈증과 혈중 총콜레스테롤 증가는 더 낮은 간세포암 발병 위험과 관련이 있었다. 이러한 반대되는 관계는 대만인 코호트로 시행된 이전 역학연구에서도 비슷하게 보고된 바 있다.[28] 본 연구에서는 고지혈증 진단명과 함께 실제 고지혈증 약을 복용하는 것이 확인된 경우만 고지혈증으로 정의하였는데, 스타틴 계열 약제가 간세포암 발병 위험을 낮춘다는 보고가 있어 어느 정도 본 연구 결과를 뒷받침하는 것으로 보인다.[25,29] 그러나 높은 총콜레스텔롤 수치가 고지혈증 약제 복용과 별개로 독립적 인자였기에,[28] 더 완전한 설명을 위해서는 관련하여 추가연구가 필요하다.

조현병 등 정신질환과 더 낮은 간세포암 발병위험 간 관계를 정확히 해석하는 것은, 본 연구에서 사용한 데이터셋에서 이러한 질환들, 즉 조현병, 알콜 등 약물사용에 의한 정신질환, 망상증이 한 군으로 묶여있기 때문에 제한이 있다. 그러나, 알콜 등 사용에 의한 정신질환의 경우 간세포암 발병위험을 증가시키는 쪽으로 작용하였을 것이기에, 본 연구 결과에서 보이는 간세포암 발병위험 감소는 조형병과 망상증과의 관계에 의한 것이라 해석하는 것이 합리적일 수 있겠다. 특히, 조현병의 경우는 메타분석에서 간세포암 발병위험을 낮추는 관련성이 보고된 바 있다.[32] 일부 연구자들은 이러한 조현병의 보호작용이 종양억제자(tumor supressor) 유전자의 발현과 조현병과의 관계로 설명하기도 하였다.[33]

본 연구의 주된 제한점은 본 모델이 한 인종, 한 국가를 대상으로 개발되었으며, 외부 검증이 이루어지지 않았다는 점일 것이다. 따라서 본 모델이 다른 대상군을 대상으로도 일반화가 잘 되어 비슷한 성능을 보일지는 확실하게 알 수 없다. 그러나, 우리는 본 연구에서 수행한 방법, 다시 말해 기계학습으로 청구자료 및 검진결과를 이용하여 중요 질환의 발병을 예측하는 것이 비슷한 여러 코호트에서 유용하게 사용될 수 있다고 믿는다. 국가코호트를 이용하는 연구였기 때문에 진정한 의미에서의 외부 데이터셋에서 검증을 하는 것은 사실상 불가능하였다. 이러한 점을 최대한 극복하고자 데이터셋을 무작위로 나누는 것이 아니라, 시간순서 대로 두 군으로 나누었다. 이렇게 시간적으로 나누는 것이 무작위로 나누는 것보다 덜 낙관적으로 모델 성능을 예측하게 된다. 실제 본 연구진은 전체 코호트를 무작위로 학습군:테스트군을 7:3으로 나누어서 분석을 시행하기도 하였는데, 이렇게 할 경우 학습군과 테스트군 간 대부분의 변수들의 특성이 유의한 차이가 없었다(부록표 5). 본 연구에 최종적으로 사용한 연구코호트에서는 학습군과 테스트

군 간 대부분의 변수들이 유의한 차이를 보인 것과 대조적이다. 학습군과 테스트군이 균질한 경우 예측모델의 재현성(reproducibility)를 주로 보게 되는 반면, 두 군이 서로 다를수록 모델의 조금 다른 상황에서도 잘 적용이 될 수 있는지를 보게 된다. 실전에서는 미래의 코호트는 현재와 계속 달라질 것이기 때문에 후자도 매우 중요하다고 할 수 있으며, 본 연구에서는 시간 기준으로 두 군을 나눔으로써 이러한 점을 충분히 고려할 수 있었다고 믿는다.

또 한 가지 제한점은 오래 전 코호트를 대상으로 모델을 구축하였는데 15년 전과 지금은 사람들의 위험인자 상태가 많이 다를 수 있어 시간적으로 보았을 때도 일반화가 잘 안될 수 있다는 점이다. 그러나 본 예측모델의 특성 상 10년 가까운 관찰기간이 필요했었다. 추후 더 최근 코호트를 이용하여 재학습 및 재검증이 이루어져야 한다고 생각한다.

결론적으로, 본 연구진은 이번 연구를 통하여 기계학습을 포함한 인공지능 알고리즘으로 국민건강보험공단 청구자료 및 국가검진 결과를 이용하여 검진 수진자의 간세포암 발병 위험을 정확하게 예측할 수 있는 예측모델을 만들고 검증하여 보았다. 추후 이러한 노력들이 실제 적용되어, 국가건강검진 수진자들이 일반적인 건강상태 정보에 더하여 암과 같은 중요한 질환의 발병 위험과 그 원인이 되는 인자를 보고받고 이에 따라 예방을 위해 노력할 수 있도록 돕는 시스템이 갖추어질 수 있기를 희망한다.

참고문헌

참고문헌

1. Liu Z, Jiang Y, Yuan H, Fang Q, Cai N, Suo C, et al. The trends in incidence of primary liver cancer caused by specific etiologies: results from the Global Burden of Disease Study 2016 and implications for liver cancer prevention. J Hepatol. 2018;70:674-83.
2. Mittal S, El-Serag HB. Epidemiology of Hepatocellular Carcinoma. J Clin Gastroenterol. 2013;47;S2-6.
3. Kim BH, Park J-W. Epidemiology of liver cancer in South Korea. Clin Mol Hepatology. 2018;24:1-9.
4. Kim S, Kim M-S, You S-H, Jung S-Y. Conducting and Reporting a Clinical Research Using Korean Healthcare Claims Database. Korean J Fam Medicine. 2020;41:146-52.
5. Hsu Y-C, Yip TC-F, Ho HJ, Wong VW-S, Huang Y-T, El-Serag HB, et al. Development of a scoring system to predict hepatocellular carcinoma in Asians on antivirals for chronic hepatitis B. J Hepatol. 2018;69:278-85.
6. El-Serag HB, Kanwal F, Davila JA, Kramer J, Richardson P. A New Laboratory-Based Algorithm to Predict Development of Hepatocellular Carcinoma in Patients With Hepatitis C and Cirrhosis. Gastroenterology. 2014;146:1249-1255.e1.
7. Kuang S-Y, Jackson PE, Wang J-B, Lu P-X, Muñoz A, Qian G-S, et al. Specific mutations of hepatitis B virus in plasma predict liver cancer development. P Natl Acad Sci Usa. 2004;101:3575-80.
8. Yang H-I, Yuen M-F, Chan HL-Y, Han K-H, Chen P-J, Kim D-Y, et al. Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): development and validation of a predictive score. Lancet Oncol. 2011;12:568-74.
9. Ripoll C, Groszmann RJ, Garcia-Tsao G, Bosch J, Grace N, Burroughs A, et al. Hepatic venous pressure gradient predicts development of hepatocellular carcinoma independently of severity of cirrhosis. J Hepatol. 2009;50:923-8.

10. Wong VW, Yu J, Cheng AS, Wong GL, Chan H, Chu ES, et al. High serum interleukin-6 level predicts future hepatocellular carcinoma development in patients with chronic hepatitis B. Int J Cancer. 2009;124:2766-70.

11. Kubota N, Fujiwara N, Hoshida Y. Clinical and Molecular Prediction of Hepatocellular Carcinoma Risk. J Clin Medicine. 2020;9:3843.

12. Lee HW, Ahn SH. Prediction models of hepatocellular carcinoma development in chronic hepatitis B patients. World J Gastroentero. 2016;22:8314-21.

13. Seong SC, Kim Y-Y, Park SK, Khang YH, Kim HC, Park JH, et al. Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. Bmj Open. 2017;7:e016640.

14. Ahn E. Introducing big data analysis using data from National Health Insurance Service. Korean J Anesthesiol. 2020;73:205-11.

15. KCD-6: Korean Standard Classification of Diseases and Causes of Death. https://koicd.kr/kcd/kcd.do?degree=06. Accessed 10 Mar 2021.

16. ICD-10: international statistical classification of diseases and related health problems : tenth revision, 2nd ed. https://apps.who.int/iris/handle/10665/42980. Accessed 10 Mar 2021.

17. Choi E-K. Cardiovascular Research Using the Korean National Health Information Database. Korean Circ J. 2019;50:754.

18. Paulino ÂDC, Guimarães LNF, Shiguemori EH. Hybrid adaptive computational intelligence-based multisensor data fusion applied to real-time UAV autonomous navigation. Inteligencia Artif. 2019;22:162-95.

19. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019;112:103375.

20. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. J Am Stat Assoc. 2012;94:496-509.

21. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. Biostatistics. 2014;15:757-73.

22. Michikawa T, Inoue M, Sawada N, Iwasaki M, Tanaka Y, Shimazu T, et al. Development of a prediction model for 10-year risk of hepatocellular carcinoma in middle-aged

Japanese: The Japan Public Health Center-based Prospective Study Cohort II. Prev Med. 2012;55:137-43.

23. Wen C-P, Lin J, Yang YC, Tsai MK, Tsao CK, Etzel C, et al. Hepatocellular Carcinoma Risk Prediction Model for the General Population: The Predictive Power of Transaminases. Jnci J National Cancer Inst. 2012;104:1599-611.

24. Sinn DH, Kang D, Cho SJ, Paik SW, Guallar E, Cho J, et al. Risk of hepatocellular carcinoma in individuals without traditional risk factors: development and validation of a novel risk score. Int J Epidemiol. 2020;49:1562-71.

25. McGlynn KA, Petrick JL, London WT. Global Epidemiology of Hepatocellular Carcinoma An Emphasis on Demographic and Regional Variability. Clin Liver Dis. 2015;19:223-38.

26. 26. Shiels MS, Cole SR, Kirk GD, Poole C. A Meta-Analysis of the Incidence of Non-AIDS Cancers in HIV-Infected Individuals. Jaids J Acquir Immune Defic Syndromes. 2009;52:611-22.

27. Niemelä O. Biomarker-Based Approaches for Assessing Alcohol Use Disorders. Int J Environ Res Pu. 2016;13:166.

28. Chiang C, Lee L, Hung S, Lin W, Hung H, Yang W, et al. Opposite association between diabetes, dyslipidemia, and hepatocellular carcinoma mortality in the middle-aged and elderly. Hepatology. 2014;59:2207-15.

29. German MN, Lutz MK, Pickhardt PJ, Bruce RJ, Said A. Statin Use is Protective Against Hepatocellular Carcinoma in Patients With Nonalcoholic Fatty Liver Disease. J Clin Gastroenterol. 2020;54:733-40.

30. Yu M-W, Chang H-C, Liaw Y-F, Lin S-M, Lee S-D, Liu C-J, et al. Familial Risk of Hepatocellular Carcinoma Among Chronic Hepatitis B Carriers and Their Relatives. Jnci J National Cancer Inst. 2000;92:1159-64.

31. Hassan MM, Spitz MR, Thomas MB, Curley SA, Patt YZ, Vauthey J-N, et al. The association of family history of liver cancer with hepatocellular carcinoma: A case-control study in the United States. J Hepatol. 2009;50:334-41.

32. Xu D, Chen G, Kong L, Zhang W, Hu L, Chen C, et al. Lower risk of liver cancer in patients with schizophrenia: a systematic review and meta-analysis of cohort studies. Oncotarget. 2017;8:102328-35.

33. Zhuo C, Wang D, Zhou C, Chen C, Li J, Tian H, et al. Double-Edged Sword of Tumour Suppressor Genes in Schizophrenia. Front Mol Neurosci. 2019;12:1.

부 록

부록

부록표 1. 기저질환의 조작적 정의

진단명	조작적 정의
간세포암종	C22.0 진단코드로 입원
간세포암 외 다른 암종	C22.0외 다른 C 진단코드로 입원
당뇨병	E11-14 + 복약
고지혈증	E78 + 복약
고혈압	(I10-13) + (입원 혹은 외래 2번 이상 방문)
만성간염바이러스감염	B15-19
HIV 바이러스	B_
조현병, 망상증, 혹은 향정신성 약물 사용에 의한 정신질환	F_
만성간질환	K72-74, K70.2-70.4
뇌졸중	I60-64 + (입원 혹은 응급실 내원) + 영상검사
허혈성 심질환	I20-25 + (입원 혹은 1개월 내 심질환으로 사망 혹은 4회 이상 외래 방문)
심방세동	I48 +(입원 혹은 2회 이상 외래 방문)
소화기계 세균성 감염	A01-09
기타 세균성 감염	A20-99
기타 바이러스성 감염	B20-34
진균 감염	B35-49
소화기계 양성종양	D12-13
갑상선 양성종양	D34
기타 양성종양	D_
혈액이상질환	D5-8
갑상선기능저하증	E00-03
기타 갑상선질환	E04-07
기타 내분비질환	E2-3
기타 대사성질환	E79, E8, E9
기분장애	F3
기질적 원인에 의한 정신질환	F0
두통	G43-44
수면장애	G47
신경증	G5-6

진단명	조작적 정의
눈질환	H0-5
귀질환	H6-9
말초혈관질환	I73
정맥류	I83
치핵	I84
울혈성 심부전	I50
만성폐쇄성폐질환	J41-44
천식	J45-46
기관지확장증	J47
외부원인에 의한 폐질환	J60-67
알레르기 비염	J30
만성 비염	J31
부비동염	J32
비강내 용종	J33
인후질환	J35-38
식도염	K20-21, K22.1, K22.7
위염	K25-29
염증성 장질환	K50-51
과민성 장질환	K58
항문질환	K60-61
알콜성 지방간	K70.0, K70.1
독성 간질환	K71
비알콜성 지방간	K76.0
담낭염	K81
기타 담낭질환	K82
담도계 질환	K83
급성췌장염	K85
기타 췌장질환	K86
피부질환	모든 L 코드
근골격계 질환	모든 M 코드
만성 신질환	N18-19 + (입원 혹은 2회 이상 외래방문)
비뇨기계 결석	N20-22
기타 비뇨기질환	N_

부록표 2. 콕스 비례-위험 회귀모형에 의한 간세포암 발병위험의 variance inflation factor (VIF)

변수	VIF	변수	VIF
AST	7.226	기타 세균 감염증	1.134
ALT	6.256	두통	1.133
이완기 혈압	2.590	심질환 과거력	1.131
수축기 혈압	2.536	갑상선 양성종양	1.129
도시 거주	2.220	인후질환	1.128
대도시 거주	2.183	소화기계 양성종양	1.127
GGT	1.600	요당	1.125
금식혈당	1.538	기관지확장증	1.124
상위 소득분위	1.503	조현병 등	1.124
성별	1.475	진균증	1.123
헤모글로빈	1.431	만성비염	1.123
중간 소득분위	1.428	만성간질환	1.119
과체중	1.234	기분장애	1.119
운동력	1.230	부비동염	1.118
간질환의 가족력	1.220	요단백	1.117
비만	1.211	당뇨병 가족력	1.114
나이	1.207	천식	1.114
혈중 콜레스테롤	1.203	기타 대사성질환	1.113
정맥류	1.188	소화기계 세균 감염증	1.112
혈액이상질환	1.185	외부 원인에 의한 폐질환	1.109
뇌졸중	1.183	담도계 질환	1.107
말초혈관질환	1.181	흡연력	1.107
눈질환	1.181	기질적 원인에 의한 정신질환	1.107
고혈압	1.179	울혈성 심부전증	1.104
혈뇨	1.177	수면장애	1.103
허혈성 심질환	1.176	HIV 감염증	1.103
뇌졸중 과거력	1.172	과민성 장질환	1.099
만성간염바이러스감염	1.169	알콜성 지방간	1.094
만성폐쇄성폐질환	1.169	항문질환	1.092
음주력	1.168	기타 담장질환	1.087
비강 용종	1.165	기타 비뇨기질환	1.085

변수	VIF	변수	VIF
고혈압 가족력	1.164	심방세동	1.074
기타 바이러스감염	1.161	기타 췌장질환	1.066
암 가족력	1.161	근골격계 질환	1.064
고지혈증	1.158	식도염	1.061
갑상선질환	1.156	피부질환	1.053
치핵	1.152	담낭염	1.038
요 pH	1.146	염증성 장질환	1.037
알레르기성 비염	1.145	급성췌장염	1.036
귀질환	1.145	비알콜성 지방간	1.025
당뇨병	1.145	독성 간질환	1.020
신경증	1.140	만성 신질환	1.016
기타 내분비질환	1.136	비뇨기계 결석	1.012

AST = Aspartate transaminase, ALT = alanine aminotransferase, GGT = gamma-glutamyl transpeptidase.

부록표 3. 학습군과 테스트군에 속하는 연구대상자들의 특성(모든 기저질환 포함)

변수		학습군	테스트군	p-value	전체
인구사회학적 특성					
나이		54.3 (9.28)	57.7 (9.6)	<0.001	55.0 (9.45)
성별	여성	42.2% (140035/331694)	55.5% (47546/85652)	<0.001	44.9% (187581/417346)
	남성	57.8% (191659/331694)	44.5% (38106/85652)		55.1% (229765/417346)
소득수준	<30%	28.5% (94614/331694)	30% (25732/85652)	<0.001	28.8% (120346/417346)
	30-80%	34.9% (115713/331694)	40.5% (34656/85652)		36% (150369/417346)
	>80%	36.6% (121367/331694)	29.5% (25264/85652)		35.1% (146631/417346)
신체계측					
체질량 지수	정상	66.2% (219447/331529)	64.2% (54942/85617)	0.213	65.8% (274419/417217)
	과체중	31.3% (103922/331529)	32.6% (27933/85617)		31.6% (131855/417217)
	비만	2.5% (8130/331529)	3.2% (2742/85617)		2.6% (10943/417217)
혈압	수축기	126.59 (17.19)	126.54 (17.17)	0.33	126.58 (17.18)
	이완기	79.18 (11.15)	78.37 (10.82)	<0.001	79.02 (11.09)
혈액검사					
AST (IU/L)		26.55 (15.92)	26.47 (17.37)	0.25	26.53 (16.23)
ALT (IU/L)		25.52 (19.45)	24.9 (20.63)	<0.001	25.39 (19.7)
GGT (IU/L)		37.7 (51.85)	36.58 (53.85)	<0.001	37.47 (52.27)
총콜레스테롤 (mg/dL)		198.32 (36.84)	199.45 (37.84)	<0.001	198.55 (37.05)
금식혈당 (mg/dL)		97.87 (28.9)	99.53 (28.12)	<0.001	98.21 (28.75)
혈색소 (g/dL)		13.94 (1.49)	13.65 (1.51)	<0.001	13.89 (1.5)
생활습관					
흡연(팩-년)		5.96 (11.53)	4.6 (11.11)	<0.001	5.68 (11.46)
음주(ml/주)		8.7 (18.7)	7.7 (19.1)	<0.001	8.5 (18.8)
운동	거의 안함	50.1% (162668/324506)	56.4% (46716/82888)	<0.001	51.4% (209384/407394)

변수	학습군	테스트군	p-value	전체
주1-2회	26.9% (87443/324506)	22.2% (18437/82888)		26% (105880/407394)
주3-4회	12.1% (39341/324506)	10.4% (8647/82888)		11.8% (47988/407394)
주5-6회	3.2% (10384/324506)	3.1% (2587/82888)		3.2% (12971/407394)
거의 매일	7.6% (24670/324506)	7.8% (6501/82888)		7.7% (31171/407394)
가족력				
간질환	2.81% (8563/305086)	2.64% (2040/77356)	0.01	2.77% (10603/382442)
고혈압	9.16% (28072/306539)	9.69% (7548/77870)	<0.001	9.27% (35620/384409)
뇌졸중	5.47% (16738/305817)	5.55% (4310/77600)	0.38	5.49% (21048/383417)
심장질환	2.39% (7278/305106)	2.57% (1992/77365)	<0.001	2.42% (9270/382471)
당뇨병	6.45% (19729/306048)	6.8% (5280/77653)	<0.001	6.52% (25009/383701)
암	13.14% (40389/307443)	13.45% (10498/78065)	0.02	13.2% (50887/385508)
기저질환				
당뇨병	6.07% (20139/331694)	8.83% (7565/85652)	<0.001	6.64% (27704/417346)
고지혈증	5.95% (19725/331694)	10.45% (8950/85652)	<0.001	6.87% (28675/417346)
고혈압	19.15% (63511/331694)	28.11% (24073/85652)	<0.001	20.99% (87584/417346)
만성간염바이러스 감염	2.57% (8530/331694)	3.71% (3176/85652)	<0.001	2.8% (11706/417346)
HIV 감염	6.73% (22314/331694)	10.6% (9079/85652)	<0.001	7.52% (31393/417346)
조현병 등 정신질환	15.59% (51714/331694)	25.57% (21905/85652)	<0.001	17.64% (73619/417346)
만성간질환	5.4% (17927/331694)	8.21% (7033/85652)	<0.001	5.98% (24960/417346)

변수	학습군	테스트군	p-value	전체
뇌졸중	0.02% (72/331694)	0.08% (67/85652)	<0.001	0.03% (139/417346)
허혈성심질환	2.67% (8859/331694)	4.72% (4046/85652)	<0.001	3.09% (12905/417346)
심방세동	0.41% (1351/331694)	0.69% (595/85652)	<0.001	0.47% (1946/417346)
소화기계 세균 감염	11.59% (38452/331694)	19.17% (16422/85652)	<0.001	13.15% (54874/417346)
기타 세균 감염	4.06% (13474/331694)	6.51% (5577/85652)	<0.001	4.56% (19051/417346)
기타 바이러스 감염	3.57% (11835/331694)	5.08% (4354/85652)	<0.001	3.88% (16189/417346)
진균 감염	21.21% (70368/331694)	24.1% (20651/85652)	<0.001	21.8% (91019/417346)
소화기계 양성종양	1.5% (4979/331694)	3.16% (2703/85652)	<0.001	1.84% (7682/417346)
갑상선 양성종양	0.42% (1379/331694)	0.9% (768/85652)	<0.001	0.51% (2147/417346)
기타 양성종양	3.11% (10323/331694)	5.66% (4847/85652)	<0.001	3.63% (15170/417346)
혈액이상질환	3.82% (12674/331694)	7.1% (6081/85652)	<0.001	4.49% (18755/417346)
갑상선질환	4.2% (13928/331694)	7.8% (6686/85652)	<0.001	4.9% (20614/417346)
기타 내분비질환	0.41% (1357/331694)	0.77% (661/85652)	<0.001	0.48% (2018/417346)
기타대사질환	1.01% (3341/331694)	2.07% (1775/85652)	<0.001	1.23% (5116/417346)
기분장애	4.48% (14854/331694)	7.72% (6613/85652)	<0.001	5.14% (21467/417346)
기질적원인에 의한 정신질환	0.23% (752/331694)	0.64% (550/85652)	<0.001	0.31% (1302/417346)
두통	10.89% (36130/331694)	18.79% (16091/85652)	<0.001	12.51% (52221/417346)
수명장애	2.39% (7926/331694)	4.88% (4184/85652)	<0.001	2.9% (12110/417346)

변수	학습군	테스트군	p-value	전체
신경증	8.58% (28464/331694)	16.23% (13905/85652)	<0.001	10.15% (42369/417346)
안질환	45.05% (149415/331694)	60.82% (52093/85652)	<0.001	48.28% (201508/417346)
귀질환	20.85% (69154/331694)	25% (21413/85652)	<0.001	21.7% (90567/417346)
말초혈관질환	3.06% (10135/331694)	6.95% (5956/85652)	<0.001	3.86% (16091/417346)
정맥류	0.54% (1792/331694)	1.13% (972/85652)	<0.001	0.66% (2764/417346)
치핵	6.11% (20267/331694)	8.87% (7594/85652)	<0.001	6.68% (27861/417346)
울혈성 심부전증	1.3% (4299/331694)	2.42% (2075/85652)	<0.001	1.53% (6374/417346)
만성폐쇄성폐질환	11.08% (36746/331694)	18.52% (15867/85652)	<0.001	12.61% (52613/417346)
천식	10.78% (35767/331694)	18.2% (15589/85652)	<0.001	12.31% (51356/417346)
기관지확장증	0.91% (3005/331694)	1.59% (1359/85652)	<0.001	1.05% (4364/417346)
외부 원인에 의한 폐질환	0.06% (198/331694)	0.11% (96/85652)	<0.001	0.07% (294/417346)
알레르기성 비염	29.9% (99186/331694)	44.23% (37886/85652)	<0.001	32.84% (137072/417346)
만성비염	10.67% (35389/331694)	14.77% (12649/85652)	<0.001	11.51% (48038/417346)
부비동염	7.14% (23693/331694)	11.17% (9571/85652)	<0.001	7.97% (33264/417346)
비강내 용종	0.79% (2622/331694)	1.14% (977/85652)	<0.001	0.86% (3599/417346)
인후질환	8.93% (29611/331694)	14.62% (12519/85652)	<0.001	10.09% (42130/417346)
식도염	14.14% (46891/331694)	23.67% (20278/85652)	<0.001	16.09% (67169/417346)
위염	56.52% (187485/331694)	74.41% (63733/85652)	<0.001	60.19% (251218/417346)

변수	학습군	테스트군	p-value	전체
염증성 장질환	0.79% (2621/331694)	1.13% (969/85652)	<0.001	0.86% (3590/417346)
과민성 장질환	17.09% (56680/331694)	26.26% (22495/85652)	<0.001	18.97% (79175/417346)
항문질환	1.46% (4853/331694)	2% (1717/85652)	<0.001	1.57% (6570/417346)
알콜성 지방간	1.95% (6480/331694)	2.75% (2353/85652)	<0.001	2.12% (8833/417346)
독성 간질환	0.88% (2908/331694)	1.4% (1195/85652)	<0.001	0.98% (4103/417346)
알콜성 지방간	3.01% (9996/331694)	5.07% (4345/85652)	<0.001	3.44% (14341/417346)
담낭염	0.26% (872/331694)	0.45% (383/85652)	<0.001	0.3% (1255/417346)
기타 담낭질환	0.11% (366/331694)	0.24% (208/85652)	<0.001	0.14% (574/417346)
담도계질환	0.13% (417/331694)	0.26% (222/85652)	<0.001	0.15% (639/417346)
급성췌장염	0.26% (877/331694)	0.47% (402/85652)	<0.001	0.31% (1279/417346)
기타 췌장질환	0.2% (679/331694)	0.42% (364/85652)	<0.001	0.25% (1043/417346)
피부질환	48.68% (161477/331694)	66% (56527/85652)	<0.001	52.24% (218004/417346)
근골격계 질환	65.09% (215916/331694)	83.78% (71762/85652)	<0.001	68.93% (287678/417346)
비뇨기계 결석	2.24% (7429/331694)	3.4% (2915/85652)	<0.001	2.48% (10344/417346)
기타 비뇨기질환	20.75% (68842/331694)	31.58% (27048/85652)	<0.001	22.98% (95890/417346)

부록표 4. 부트스트랩으로 1000번 반복하여 Cox 비례-위험 회귀분석을 반복하였을 때, 각 변수가 간세포암발병과 유의한 관계를 보였던 빈도

변수	빈도	변수	빈도
나이	100% (1,000/1,000)	성별	100% (1,000/1,000)
만성간질환	100% (1,000/1,000)	GGT	100% (1,000/1,000)
만성간질환 가족력	100% (1,000/1,000)	ALT	100% (1,000/1,000)
만성간염바이러스감염	100% (1,000/1,000)	총콜레스테롤	100% (1,000/1,000)
HIV 감염	100% (1,000/1,000)	고지혈증	99.1% (991/1,000)
당뇨병	99.1% (991/1,000)	조현병 등 정신질환	98.2% (982/1,000)
소득수준	88.8% (888/1,000)	비만	86.5% (865/1,000)
진균증	83.5% (835/1,000)	음주습관	81.2% (812/1,000)
기타 비뇨기계질환	79.8% (798/1,000)	식도염	76.9% (769/1,000)
요 단백	75.6% (756/1,000)	만성비염	74.7% (747/1,000)
요 당	73.0% (730/1,000)	치핵	67.3% (673/1,000)
갑상선질환	64.2% (642/1,000)	과민성 장질환	61.9% (619/1,000)
비알콜성 지방간	55.5% (555/1,000)	알콜성 지방간	51.6% (516/1,000)
당뇨병가족력	46.2% (462/1,000)	귀질환	45.9% (459/1,000)
두통	41.8% (418/1,000)	알레르기성 비염	39.4% (394/1,000)
비뇨기계 결석	38.4% (384/1,000)	담낭염	37.4% (374/1,000)
고혈압	28.6% (286/1,000)	운동습관	28.4% (284/1,000)
기타 세균 감염	28.0% (280/1,000)	흡연	22.4% (224/1,000)
소화기계 세균 감염	21.1% (211/1,000)	기타 양성종양	20.0% (200/1,000)

변수	빈도	변수	빈도
기타 담낭질환	19.1% (191/1,000)	외부 원인에 의한 폐질환	17.7% (177/1,000)
피부질환	16.6% (166/1,000)	혈색소	13.4% (134/1,000)
기분장애	11.9% (119/1,000)	금식 혈당	11.8% (118/1,000)
심질환 가족력	11.3% (113/1,000)	위염	10.4% (104/1,000)
거주지역	10.8% (108/1,000)	안질환	10.1% (101/1,000)
소화기계 양성종양	9.9% (99/1,000)	급성췌장염	9.7% (97/1,000)
인후질환	9.5% (95/1,000)	기타 바이러스 감염	9.4% (94/1,000)
정맥류	8.8% (88/1,000)	기타 세균 감염	8.1% (81/1,000)
독성 간질환	7.8% (78/1,000)	만성폐쇄성폐질환	7.8% (78/1,000)
기타 담도계질환	7.3% (73/1,000)	신경증	7.2% (72/1,000)
요 pH	7.0% (70/1,000)	근골격계질환	6.5% (65/1,000)
요 잠혈	6.1% (61/1,000)	항문질환	5.8% (58/1,000)
비강내 용종	5.5% (55/1,000)	뇌졸중 가족력	5.3% (53/1,000)
기타 대사성질환	5.3% (53/1,000)	암 가족력	5.2% (52/1,000)
심방세동	4.8% (48/1,000)	혈액이상질환	4.6% (46/1,000)
장애	4.3% (43/1,000)	기관지확장증	3.7% (37/1,000)
부비동염	3.2% (32/1,000)	기타 췌장질환	2.8% (28/1,000)
천식	2.6% (26/1,000)	수면장애	2.6% (26/1,000)
허혈성심질환	2.6% (26/1,000)		

AST = Aspartate transaminase, ALT = alanine aminotransferase, GGT = gamma-glutamyl transpeptidase.

부록표 5. 학습군과 테스트군을 무작위로 나누었을 때 두 군의 연구대상자들 특성(모든 기저질환 포함)

변수		학습군	테스트군	p-value	전체
인구사회학적 특성					
나이		54.3 (9.28)	54.27 (9.28)	0.43	54.29 (9.28)
성별	여성	42.2% (98048/232186)	42.2% (41987/99508)	0.86	42.2% (140035/331694)
	남성	57.8% (134138/232186)	57.8% (57521/99508)		57.8% (191659/331694)
소득수준	<30%	16.7% (38808/232186)	16.5% (16371/995038)	0.15	16.6% (55179/331694)
	30-80%	28.2% (65516/232186)	28.2% (28066/99508)		28.2% (93582/331694)
	>80%	55.1% (127862/232186)	55.3% (55071/99508)		55.2% (182933/331694)
신체계측					
체질량 지수	정상	23.93 (2.88)	23.94 (2.89)	0.41	23.93 (2.89)
	과체중	66.2% (153721/232070)	66.1% (65756/99459)	0.9	66.2% (219447/331529)
	비만	31.3% (72675/323070)	31.4% (31247/99459)		31.3% (103922/331529)
혈압	수축기	2.4% (5674/232070)	2.5% (2456/99459)		2.5% (8130/331529)
	이완기	126.57 (17.17)	126.63 (17.23)	0.39	126.59 (17.19)
혈액검사					
AST (IU/L)		26.56 (16)	26.52 (15.72)	0.42	26.55 (15.92)
ALT (IU/L)		25.52 (19.56)	25.50 (19.18)	0.77	25.52 (19.45)
GGT (IU/L)		37.73 (52.2)	37.64 (51.03)	0.66	37.7 (51.85)
총콜레스테롤 (mg/dL)		198.37 (36.86)	198.2 (36.8)	0.2	198.32 (36.84)
금식혈당 (mg/dL)		97.89 (29.18)	97.81 (28.24)	0.42	97.87 (28.9)
혈색소 (g/dL)		13.95 (1.49)	13.94 (1.49)	0.73	13.94 (1.49)
생활습관					
흡연(팩-년)		5.98 (11.56)	5.92 (11.46)	0.18	5.96 (11.53)
음주(ml/주)		56.3% (128639/2283638)	56.2% (590593/9/8/35)	0.48	56.3% (183692/326241)

변수		학습군	테스트군	p-value	전체
운동	거의 안함	15.6% (35665/228368)	15.5% (15180/97873)		15.6% (50845/326241)
	주1-2회	17.4% (39629/228368)	17.5% (17135/97873)		17.4% (56764/326241)
	주3-4회	6.8% (15453/228368)	6.9% (6744/97873)		6.8% (22197/326241)
	주5-6회	3.9% (8982/228368)	3.8% (3761/97873)		3.9% (12743/326241)
	거의 매일	50.1% (113877/227/095)	50.1% (48791/97411)	0.59	50.1% (162668/324506)
가족력					
간질환		2.8% (5989/213528)	2.81% (2574/91558)	0.9	2.81% (8563/305086)
고혈압		9.16% (19648/214527/)	9.16% (8424/92012)	0.77	9.16% (28072/306539)
뇌졸중		5.49% (11742/214048)	5.44% (4996/91769)	0.85	5.47% (16738/305817)
심장질환		2.39% (5102/213537)	2.38% (21/76/91569)	0.85	2.39% (7278/305106)
당뇨병		6.4% (13718/214195)	6.54% (6011/91853)	0.3	6.45% (19729/306048)
암		13.1% (28189/215156)	13.22% (12200/9228/)	0.49	13.14% (40389/307443)
기저질환					
당뇨병		6.08% (14114/232186)	6.05% (6025/99508)	0.8	6.07% (20139/331694)
고지혈증		5.95% (13811/232186)	5.94% (5914/99508)	0.96	5.95% (19725/331694)
고혈압		19.08% (44298/232186)	19.31% (19213/99508)	0.13	19.15% (63511/331694)
만성간염바이러스 감염		2.55% (5915/232186)	2.63% (2615/99508)	0.16	2.57% (8530/331694)
HIV 감염		6.69% (15527/232186)	6.82% (6787/99508)	0.16	6.73% (22314/331694)
조현병 등 정신질환		15.55% (36110/232186)	15.68% (15604/99508)	0.35	15.59% (51714/331694)
만성간질환		5.77% (13397/232186)	5.17% (5149/99508)	0.15	5.59% (18546/331694)

변수	학습군	테스트군	p-value	전체
뇌졸중	0.02% (56/232186)	0.02% (16/99508)	0.19	0.02% (72/331694)
허혈성심질환	2.68% (6232/232186)	2.64% (2627/99508)	0.48	2.6/% (8859/331694)
심방세동	0.41% (952/232186)	0.4% (399/99508)	0.73	0.41% (1351/331694)
소화기계 세균 감염	11.55% (26818/232186)	11.69% (11634/99508)	0.25	11.59% (38452/331694)
기타 세균 감염	5.10% (11920/232186)	5.01% (4983/99508)	0.74	5.10% (16903/331694)
기타 바이러스 감염	6.22% (14434/232186)	6.18% (6145/99508)	0.66	6.2% (205/9/331694)
진균 감염	21.14% (49092/232186)	21.38% (21294/99508)	0.13	21.21% (70368/331694)
소화기계 양성종양	1.51% (3500/232186)	1.49% (1479/99508)	0.66	1.5% (4979/331694)
갑상선 양성종양	0.41% (942/232186)	0.44% (437/99508)	0.18	0.42% (1379/331694)
기타 양성종양	3.09% (7176/232186)	3.16% (3147/99508)	0.28	3.11% (10323/331694)
혈액이상질환	3.83% (8883/232186)	3.81% (3791/99508)	0.83	3.82% (126/4/331694)
갑상선질환	4.2% (7176/232186)	4.19% (4168/99508)	0.85	4.2% (13928/331694)
기타 내분비질환	0.42% (964/232186)	0.39% (393/995038)	0.42	0.41% (135//331694)
기타대사질환	1.01% (2334/232186)	1.01% (1007/99508)	0.87	1.01% (3341/331694)
기분장애	4.47% (10370/232186)	4.51% (4484/99508)	0.62	4.48% (14854/331694)
기질적원인에 의한 정신질환	0.23% (531/232186)	0.22% (221/99508)	0.74	0.23% (752/331694)
두통	10.93% (25379/232186)	10.8% (10751/99508)	0.29	10.89% (36130/331694)
수명장애	2.36% (5479/232186)	2.46% (2447/9950)	0.09	2.39% (7926/331694)
신경증	3.96% (19878/232186)	8.63% (8586/99508)	0.53	8.58% (28464/331694)

변수	학습군	테스트군	p-value	전체
안질환	45.05% (104610/232186)	45.03% (44805/99508)	0.89	45.05% (149415/331694)
귀질환	20.89% (48513/232186)	20.74% (20641/99508)	0.33	20.85% (69154/331694)
말초혈관질환	3.06% (7114/232186)	3.04% (3021/99508)	0.68	3.06% (10135/331694)
정맥류	0.54% (1252/232186)	0.54% (540/99508)	0.92	0.54% (1792/331694)
치핵	6.13% (14237/232186)	6.06% (6030/99508)	0.43	6.11% (2026//331694)
울혈성 심부전증	1.28% (2973/232186)	1.33% (1326/99508)	0.23	1.3% (4299/331694)
만성폐쇄성폐질환	11.1% (25772/232186)	11.03% (10974/99503)	0.55	11.08% (36746/331694)
천식	10.77% (25009/232186)	10.81% (10758/99508)	0.74	10.78% (35767/331694)
기관지확장증	0.92% (2134/232186)	0.88% (871/99508)	0.23	0.91% (3005/331694)
외부 원인에 의한 폐질환	0.06% (132/232186)	0.07% (66/99508)	0.34	0.06% (198/331694)
알레르기성 비염	30.04% (69750/232186)	29.58% (29436/99508)	0.01	29.9% (99186/331694)
만성비염	10.71% (24869/232186)	10.57% (10520/99508)	0.24	10.67% (35389/331694)
부비동염	7.19% (16690/2321386)	7.04% (7003/99508)	0.12	7.14% (23693/331694)
비강내 용종	0.79% (1843/232186)	0.78% (779/99508)	0.76	0.79% (2622/331694)
인후질환	8.96% (20793/232186)	8.86% (8818/99508)	0.39	8.93% (29611/331694)
식도염	14.2% (329/0/232186)	13.99% (13921/99508)	0.11	14.14% (46891/331694)
위염	96.53% (131248/232186)	96.52% (56237/99508)	0.95	56.52% (18/485/331694)
염증성 장질환	0.78% (1822/232186)	0.8% (799/99508)	0.6	0.79% (2621/331694)
과민성 장질환	17.13% (39784/232186)	16.98% (16896/99508)	0.28	17.09% (56680/331694)

변수	학습군	테스트군	p-value	전체
항문질환	1.49% (3449/232186)	1.41% (1404/99508)	0.1	1.46% (4853/331694)
알콜성 지방간	3.2% (7427/232186)	3.21% (3192/99508)	0.9	3.2% (10619/3316094)
독성 간질환	0.88% (2042/232186)	0.87% (866/99508)	0.81	0.88% (2908/331694)
알콜성 지방간	9.23% (21432/232186)	9.11% (9065/99508)	0.27	9.19% (30497/331694)
담낭염	0.26% (598/232186)	0.28% (274/99508)	0.38	0.26% (8/2/331694)
기타 담낭질환	0.11% (259/232186)	0.11% (107/99508)	0.79	0.11% (366/331694)
담도계질환	0.13% (293/232186)	0.12% (124/99508)	0.95	0.13% (417/331694)
급성췌장염	0.26% (612/232186)	0.27% (265/99508)	0.92	0.26% (877/331694)
기타 췌장질환	0.21% (480/232186)	0.2% (199/99508)	0.72	0.2% (679/331694)
피부질환	48.73% (113151/232186)	43.96% (48326/99508)	0.38	43.68% (161477/331694)
근골격계 질환	65.19% (1513 /0/232186)	64.87% (64546/99508)	0.07	65.09% (215916/331694)
비뇨기계 결석	2.29% (5220/232186)	2.22% (2209/99508)	0.62	2.24% (7429/331694)
기타 비뇨기질환	20.71% (48077/232186)	20.87% (20765/99508)	0.3	20.75% (68842/331694)

연구보고서 2020-20-031

의료이용 행태를 분석하여 중증질환 발생 위험을 예측하는 인공지능 알고리듬의 개발 및 검증: 대규모 한국 코호트 연구

발 행 일	2022년 2월 28일
발 행 인	김성우
편 집 인	이천균
발 행 처	국민건강보험 일산병원 연구소
주 소	경기도 고양시 일산동구 일산로 100(국민건강보험 일산병원)
전 화	031) 900-6977, 6985
팩 스	0303-3448-7105~7
인 쇄 처	지성프린팅 (02-2278-2490)

(우)10444 경기도 고양시 일산동구 일산로 100(백석1동 1232번지)
대표전화 1577-0013 / 팩스 031-900-0049
www.nhimc.or.kr

의료이용 행태를 분석하여 중증질환 발생 위험을 예측하는 인공지능 알고리듬의 개발 및 검증: 대규모 한국 코호트 연구

National Health Insurance Service Ilsan Hospital