JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 5, pp. 1407-1419, May. 2025

Check for updates

만성질환 예측과 상대위험도 산출을 통한 인슈어테크 적용 가능성 연구

류 범 상[1] · 성 지 민[2*]
[1]주식회사 온택트헬스 연구원
[2]연세대학교 의과대학 뇌혈관질환연구센터 연구교수

Exploring InsurTech Applications through Chronic Disease Prediction and Relative Risk Modeling

Beom-Sang Ryu[1] · Ji-Min Sung[2*]

[1]Researcher, ONTACT HEALTH Co.,Ltd., Seoul 03764, Korea
[2]Research Professor, Integrative Research Center for Cerebrovascular and Cardiovascular Diseases, Yonsei University College of Medicine, Seoul 03722, Korea

[요 약]

본 연구는 국민건강보험 표본코호트 DB의 세부 의료이용 내역을 활용해 암·뇌·심장질환 등 대표적인 3대 만성질환의 발생을 예측하는 머신러닝 모형을 구현하고, 인슈어테크(InsurTech) 관점에서 건강보험의 가격책정(Pricing)과 인수심사(Underwriting) 분야에 적용 가능성을 검토하였다. 이를 위해 데이터 기반의 변수 선택 전략을 통해 질환 발생의 주요 예측요인을 식별하고, 집단별 상대위험도를 산정하여 기존 보험산업의 표준체와 비표준체 분류체계의 세분화 가능성을 평가하였다. 분석 결과 예측 대상 질환에 따라 예측요인의 구성이 상이하고, 표준체와 간편고지체, 거절체 순으로 상대위험도가 체계적으로 높아지는 양상이 확인되었다. 이는 보건의료 데이터와 머신러닝 기법 활용이 기존 보험산업의 위험분류 체계와 충돌되지 않으면서도 고위험군 내부의 위험을 보다 세분화하여 개인 맞춤형 보험료 책정과 가입대상 확대에 기여할 수 있음을 시사한다.

[Abstract]

This study develops a machine learning model predicting the incidence of cancer, cerebrovascular disease, and cardiac diseases using medical utilization data from the National Health Insurance Service cohort database (DB). Adopting an InsurTech perspective, we evaluate the model's potential to enhance pricing and underwriting in health insurance. The data-driven variable selection identifies critical disease-specific predictors, while segment-specific relative risks are calculated to explore the refinement of traditional insurance risk classifications. Findings reveal distinct predictors by disease and systematically higher relative risks from standard to simplified-issue and declined groups, suggesting that predictive modeling using medical utilization data can refine risk stratification, thus enabling personalized pricing and an expanded underwriting.

색인어 : 질병예측모형, 인슈어테크, 보험료 세분화, 보건의료빅데이터, 머신러닝
Keyword : Disease Prediction Model, Insurtech, Premium Segmentation, Healthcare Big Data, Machine Learning

http://dx.doi.org/10.9728/dcs.2025.26.5.1407

Received 16 April 2025; **Revised** 29 April 2025
Accepted 29 April 2025

***Corresponding Author; Ji-Min Sung**

Tel: +82-2-2228-2688
E-mail: jmsung@yuhs.ac

Ⅰ. 서 론

보험은 개인의 예기치 못한 건강 문제나 재정적 손실에 대해 금전적 보상을 제공하는 기능을 한다. 이를 위해 보험사는 보험상품의 개발과정에서 다양한 데이터를 바탕으로 위험요인을 체계적으로 평가해 반영하고, 서로 다른 위험 수준을 가진 가입자들의 보험료를 합리적으로 부과할 수 있도록 설계함으로써 가입자 간 부담의 공정성과 보험재정의 안정성을 추구한다.

최근 급격한 인구구조 변화로 고령층과 유병자와 같이 상대적으로 위험 수준이 높은 계층이 빠르게 증가함에 따라 보험산업은 새로운 도전에 직면하게 되었다. 통계청 인구동향조사에 따르면 우리나라는 2024년 주민등록 인구를 기준으로 65세 이상의 고령인구 비중이 20%를 넘어서며 초고령사회로 진입했다. 또한 2023년 기준 가임 여성 1명당 합계출산율이 0.72명으로 추계하며 인구구조 변화가 급속도로 진행되는 상황이다. 이러한 인구구조 변화로 건강문제와 그에 따른 재무적 불확실성이 높은 고령인구가 증가하며 건강보험의 보장 수요는 증가할 것으로 전망되고 있다[1]. 반면, 보험산업의 관점에서 인구구조 변화는 일반 위험률을 적용하는 표준체 중심의 시장 감소를 의미한다. 또한, 상대적으로 위험 수준이 높은 계층으로의 시장 확대는 보험계약 시 예상했던 것보다 많은 보험금이 지급되는 보험위험의 가능성이 높아지는 원인이 될 수 있다. 따라서 현재의 보험산업은 시장 정체에 대응하여 가입 요건을 완화해 시장을 확대하면서도 보험위험을 적절한 수준에서 통제하여 보험료 부담의 공정성과 보험재정의 안정성을 달성해야 하는 상황에 놓여있다[2].

이러한 환경 속에서 보건의료 빅데이터와 머신러닝·AI 기술을 접목한 인슈어테크(Insurtech)가 주목받고 있다[3]. 최근 금융위원회는 인구구조와 기술변화 등 외부환경 변화에 보험산업이 능동적으로 대응하고, 새로운 성장 동력을 마련할 수 있도록 인슈어테크 활성화를 보험산업의 미래 과제 중 하나로 제시한 바 있다[4]. 이에 따라 소비자 편익과 보험 산업 혁신을 위한 지원이 본격화되고 있으며, 관련 생태계 조성이 확대되고 있다. 특히 최근에는 개인 동의에 기반 데이터 수집과 공공 보건의료 빅데이터 개방으로 보험산업이 활용할 수 있는 의료 및 건강 데이터 접근성이 개선되고 있다. 여기에 웨어러블·텔레매틱스 등 IoT 기술로 수집되는 생활 및 정보 등 더 광범위한 데이터를 활용해 다양한 가입자의 특성을 이해하고 위험 수준을 더욱 세분화된 상품 및 서비스를 제공할 기회가 증가하고 있다. 실례로 건강증진형 보험이나 건강 수준이 일정 기준을 충족하면 보험료를 할인해주는 보험상품들이 등장하고 있다. 이들 보험은 헬스케어를 접목하고 개인 건강 수준과 연동한 보험료 세분화 구조를 통해 소비자의 의료비 부담과 보험료 형평성을 동시에 개선하고, 보험사 입장에서도 보험위험을 낮춰 손해율을 관리할 수 있는 윈윈(win-win) 구조가 가능하다는 점에서 의미가 크다[5].

이러한 맥락에서 본 연구는 국민건강보험 표본코호트 DB 2.0의 세부 의료이용 내역을 활용하여 대표적 3대 만성질환인 암·뇌·심장질환 발생률 예측모형을 구현한다. 구체적으로 과거 질병의 진단, 약제 처방, 수술 및 검사 내역을 활용하여 데이터 분석 관점에서 질환 발생 위험을 예측하는 요인들을 식별한다. 이후 예측모형을 통해 연령과 성별, 표준체와 간편고지체와 같은 보험 가입유형 등 집단별 상대위험도를 산출하여 위험도 격차를 검토한다. 이를 통해 보건의료데이터와 머신러닝을 활용한 정교한 보험위험 예측 가능성을 검토하고, 인슈어테크 기술로써 보험가입 대상 확대와 보험료 세분화 측면에서의 다양한 시사점을 논의한다.

Ⅱ. 선행연구

본 연구와 관련된 선행연구는 크게 보건의료데이터를 활용해 질환 발생률 예측모형 개발 연구와 이러한 예측모형을 활용해 보험산업의 시장 확대 및 보험료 세분화 가능성을 분석한 연구들로 구분할 수 있다. 먼저, 질환발생을 예측모형과 관련된 대표적 연구사례는 심뇌혈관질환을 비롯해 주요 만성질환의 예측모형으로 QRISK3(QResearch Risk estimator version3) 연구가 있다. 해당 모형은 2016년 기준 1800만 명이 넘는 환자의 익명화된 건강 기록을 1000개 기관에서 수집한 영국의 대표적인 보건의료 데이터베이스인 QReaserch를 기반으로 한다[6]. 머신러닝 및 딥러닝 알고리즘을 활용해 특정 질환을 예측하는 모형들도 존재하며 대표적으로 유방암 질환에 대하여 생활습관이나 호르몬 수치, 과거력에 대한 설문 정보를 활용하여 발생률을 예측한 사례가 존재한다[7].

국내의 경우 국민건강보험공단의 표본코호트 및 맞춤형 데이터를 활용한 연구들이 활발하다. 예컨대 [8]에서는 질환 예측에 검진정보를 활용하여 Lasso Logistic 모형을 비롯해 랜덤포레스트 등 다양한 머신러닝 알고리즘을 활용하여 10년 이내 질환의 발생 여부, 의료비와 질병 부담 등 예측한 바 있다. 특정 인구집단에 한정해 예측 모형을 개발한 사례로 국민건강보험공단의 노인코호트 DB를 활용해 고령인구의 뇌졸중 질환을 예측하는 연구를 들 수 있다. 해당 연구는 예측을 위해 합성곱신경망을 활용하였으며 고령층의 뇌졸중의 예방 및 조기 발견을 위한 시스템으로 활용할 수 있는 모형 구조를 검증하였다[9]. 국방의료데이터 상의 건강검진 정보를 활용해 딥러닝 기반의 폐렴 진단 및 예측 연구도 존재한다. 군의료진이 환자를 효율적으로 선별하고 진단을 보조할 수 있도록, 예측모형을 군 복무자 데이터를 기반으로 설계하였다[10]. 이외에도 건강검진 및 진료내역 데이터를 기반으로 유방암 환자의 생존 예측 모형을 개발한 사례도 존재한다. 해당 연구는 건강보험 청구 데이터를 활용하여 유방암 환자의 생존을 예측하는 모형을 구현하였으며 의사결정 나무, 랜덤 포레스트 및 그래디언트 부스팅 등 다양한 머신러닝 기법을 적용하여

머신러닝 기법이 예측 결과의 정교함과 신뢰성을 높이는 데 기여했음을 보였다[11].

한편, 보험산업 적용 관점에서 보건의료데이터와 예측모형을 활용한 의미있는 연구들이 존재한다. 이들 연구는 국민건강코호트 DB의 검진정보를 활용해 다양한 질환의 발생 위험도를 평가하는 평점화 모형을 도출하고 평점을 기준으로 상대위험도를 산출하여 보험세분화 및 보험상품 적용 가능성을 검토하였다[12],[13]. 해당 연구는 신용점수와 유사한 방식을 통해 개인의 건강을 측정가능한 형태로 제시함으로써 보험산업의 건강증진형 보험상품 개발과 건강관리 서비스 제공을 위한 건강평가 기준을 제시하였다. 이와 더불어 간편고지보험과 같이 계약전 고지의무 항목을 축소해 유병자 및 고령자 대상의 시장접근성을 높이면서 정교한 위험분석과 정량화를 통해 보험료 세분화를 검토하는 연구들도 다수 보고되었다[14]. 이외에도 예측 대상을 질환에 한정하지 않고 의료 이용 규모를 예측하는 모형 개발을 시도하여 표준체와 간편고지체의 보험사고 가능성을 간접적으로 파악하고자 하는 연구 사례도 존재한다[15].

종합하면, 본 연구는 질환 발생 예측모형을 활용하여 집단간 위험도 차이를 비교·분석한다는 점에서 기존 연구와 맥을 같이 한다. 다만, 대부분의 기존 연구가 건강검진 정보 중심으로 개인 건강 수준을 반영한 데 비해, 본 연구는 진단, 처방, 수술·검사 등 실제 의료이용 기록을 폭넓게 활용하였다. 이로 인해 건강검진을 수검하지 않은 대상자도 분석에 포함할 수 있었으며, 분석 표본의 대표성과 건강 상태 반영의 정확도를 동시에 제고하였다. 결과적으로, 의료이용 기반의 예측모형은 검진정보에 대한 의존을 완화하고, 인슈어테크 기반 보험설계에 있어 보건의료 데이터 활용 가능성에 시사점을 제공한다는 점에서 의의가 있다.

Ⅲ. 연구모형

3-1 활용자료

본 연구는 연세의료원 세브란스병원 연구심의위원회에서 연구계획 승인을 받아(승인번호 4-2023-0700) 국민건강보험공단 표본 코호트 DB 2.2을 활용하였다(건강보험공단 연구관리번호 NHIS-2023-2-246). 국민건강보험공단 표본 코호트 DB는 2006년을 기준으로 전체 건강보험 가입자의 2%인 약 100만 명을 표본 추출하여 구성된 후향적 코호트 자료이다. 국민건강보험공단이 연구 목적으로 구축한 이 DB는 고도의 대표성을 갖춘 국가 보건의료자료다. 개별 코호트에 대하여 2002년부터 2019년까지의 진단, 약제의 처방, 검사 및 수술 등 의료서비스 이용에 대한 세부적인 내용을 추적 관찰 가능하며 전체 국민건강보험 가입자에 대한 통계적 대표성을 보유하고 있다. 본 연구에서는 표본 코호트 DB의 구성 항목들 중 개별 코호트의 성별과 연령 정보를 확인할 수 있는 자격 및 보험료 테이블과 출생 및 사망 테이블을 비롯해 세부 질병의 진단 및 의료행위 정보를 식별할 수 있는 명세서(20T), 진료내역(30T), 상병내역(40T), 처방전교부상세내역(60T)을 분석에 활용하였다.

3-2 연구설계

본 연구에서는 2015년 1월 1일을 기준시점으로 설정하고, 이를 중심으로 과거 5년(2010~2014년)을 관찰기간(Observation Period), 이후 5년(2015~2019년)을 추적기간(Follow-up Period)으로 정의하였다. 이는 보험 실무의 기준과 활용자료의 시계열적 제약과 선행연구를 종합적으로 고려한 결과이다. 우선, 관찰기간을 5년으로 한정한 것은 보험계약 시 피보험자에게 요구되는 입원, 수술, 질병 이력에 대한 고지의무 기간이 표준사업 방법서 기준 통상적으로 최근 5년 이내로 규정하고 있는 점을 고려 하였다. 한편, 추적기간은 보험료의 갱신주기가 통상적으로 3년, 5년, 10년 주기로 구성되기 때문에 정해진 기준이 존재하지 않으며 자료의 특성과 목적에 따라 유연하게 설정될 수 있다. 실제로 다수의 선행연구에서는 표본코호트 DB의 시계열 범위 및 초기 구간 데이터 품질을 고려하여 추적기간을 3년 또는 5년으로 설정하는 사례가 다수 존재한다[13],[15]. 본 연구에서는 기존 선행연구의 설계를 참고해 기준시점을 2015년으로 하고 추적 기간을 5년으로 설정하되 개발된 모형을 활용해 기준시점의 변경과 5년 이내의 연차별(1~4년) 추적기간 변동에 따른 예측성능을 검증하는 방식을 취하였다. 이를 통해 시간적으로 변화된 환경에서도 일관된 예측력을 유지하는 모형을 도출하고자 하였다. 이에 대한 구체적인 내용은 4.2의 예측모형 검증 과정에서 함께 제시한다.

예측 모형 구현을 위한 자료 구축은 기준시점(2015년 1월 1일)에 국민건강보험 자격을 유지하는 코호트를 분석대상으로 한정하고, 관찰기간 내에 의료이용(진단명, 약제처방, 검사·처치·수술 이력 등) 정보를 활용해 설명변수를 구성하였다.

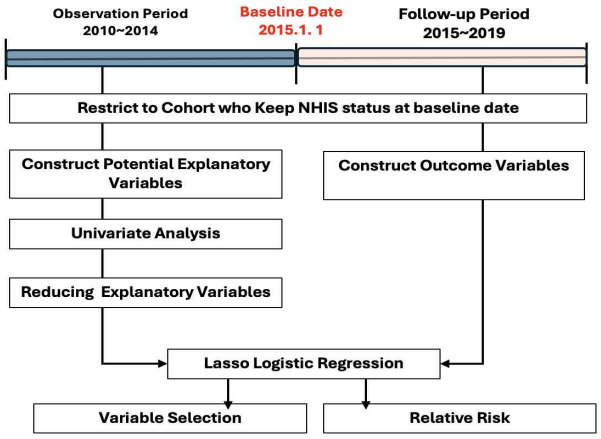

그림 1. 연구 모형
Fig. 1. Research design

예측모형의 결과변수는 추적기간동안 발생한 암(일반암, 유사암), 뇌(뇌혈관질환, 뇌출혈), 심장질환(허혈성심장질환, 급성심근경색) 총 6종의 세부 질환을 대상으로 정의하였다. 또한 질환별·성별로 예측모형을 각각 구분하여 성별 차이를 직접 고려하였다.

구축한 분석자료를 기반으로 Lasso 로지스틱 회귀 기반의 예측모형을 도출한다. 이 과정에서 Lasso 방법론을 통해 질환 발생에 유요한 예측요인으로서 설명변수를 선택하였다. 설명변수의 선택과정은 후술하지만 광범위한 의료이용정보를 활용해 발생하는 설명변수 증가에 따른 차원문제를 완화하기 위해 설명변수를 추출하고 결과변수들과의 단변량 분석을 통해 설명변수 후보군을 축소하여 모형에 반영해 최종 예측요인을 선별하는 방식을 취하였다.

예측모형으로 개인의 예측 발생률을 산출하고, 인구집단에 대한 평균값을 바탕으로 표준체 대비 간편체, 거절체의 상대위험도(Risk Ratio)를 산출하였다. 이를 통해 보험 가입 대상의 유형에 따라 상대위험도를 비교하는 분석구조를 취한다.

표 1. 질병 분류 및 진단 기준
Table 1. Disease classification and diagnostic criteria

Disease category	Assessment criteria
General cancer	All C codes(except C44, C73) D45, D46, D47.1,D47.3,D47.4,D47.5 & Special exemption code V193
Carcinoma in Situ	C44, C73, D00-D09, D37-D44, D47.0, D47.2, D47.7, D47.9, D48& Special exemption codeV193
Cerebrovascular	I60-I69
Stroke	I60-I62
Ischemic heart disease	I20-I25
Acute myocardial infarction	I21-I23

3-3 결과변수

결과변수로 예측하는 질환의 정의는 선행 연구[15]와 해당 질환을 보장하는 국내 보험상품 약관을 참고하여 작성하였으며, 표 1에 제시된 바와 같다. 암질환의 경우 일반암, 유사암으로 구분하며 뇌질환은 뇌혈관질환과 뇌출혈, 심장질환은 허혈성 심장질환과 급성 심근경색 질환을 대상으로 한다. 결과변수 구성을 위한 질환 식별은 추적기간내 대상 질환의 최초 여부를 식별해 정의하였다. 자료 상에서 각 질환의 식별을 한국표준질병사인분류(KCD) 코드를 기준으로 하였으며, 암 질환에 대해서는 본인 일부 부담금 산정 특례에 관한 기준 제 4조에서 정의하는 특정 기호(V193)를 함께 고려하였다.

3-4 분석대상

표 2는 본 연구의 분석대상자의 일반적 특성을 보여준다. 분석 대상자는 표본코호트의 자격 DB에서 성별, 출생 연도 및 사망 연도를 추출한 후, 2014년 건강보험 자격을 유지한 대상자로 한정하여 각 결과변수의 질환에 대해 구분한 분석 표본을 구축하였다. 2015년 기준 20세 이상 전체 가입 자격자는 여성 405,550명, 남성 401,271명으로 나타났다.

표 2. 성별·연령대별 보험 가입 가능 유형 분포
Table 2. Insurance eligibility types by sex and age group

	Standard group		Simplified group		Rejected group	
Age group	Female	Male	Female	Male	Female	Male
20-29	39,482	43,976	24,885	27,867	445	570
	27%	27%	11%	14%	1%	2%
30-39	27,494	43,865	48,337	36,084	987	1,454
	19%	27%	22%	18%	3%	4%
40-44	21,481	22,052	21,379	22,976	1,011	1,661
	15%	14%	10%	11%	3%	5%
45-49	19,056	17,909	21,804	23,883	1,617	2,640
	13%	11%	10%	12%	4%	7%
50-54	14,714	13,884	23,876	25,639	2,758	4,179
	10%	9%	11%	13%	8%	12%
55-59	10,509	9,317	22,446	22,813	3,813	5,242
	7%	6%	10%	11%	11%	15%
60-64	5,626	4,528	15,995	15,341	3,851	4,674
	4%	3%	7%	8%	11%	13%
65-69	3,294	2,750	12,785	11,368	4,439	4,762
	2%	2%	6%	6%	12%	13%
70+	4,571	3,343	31,750	17,587	17,145	10,908
	3%	2%	14%	9%	48%	30%
Total	307,851		426,815		72,156	

예측모형 도출 이후 상대위험도 산출과 분석을 위해 분석대상자 기반으로 보험가입유형은 표준체와 간편체, 거절체로 분류하였다. 각 유형의 정의는 다음과 같은 기준으로 하였다.

표준체: 최근 5년이내에 입원, 수술, 계속하여 7일 이상 치료 경험이 없고, 진찰 또는 검사를 통해 질병확정진단, 치료 입원, 수술, 투약 사실이 없으며, 10대 질병(암, 백혈병, 고혈압, 협심증, 심근경색, 심장판막증, 간경화증, 뇌졸중, 당뇨병, 에이즈(AIDS) 및 HIV 보균) 에 의한 진단 및 수술, 입원 사실이 없는 자

간편체: 최근 5년 이내에 질병이나 상해 사고로 입원 또는 수술을 받은 사실이 없고, 6대 질병(암, 협심증, 심근경색, 심장판막증, 간경화증, 뇌졸중)에 의한 진단 및 수술, 입원 사실이 없는 자

거절체: 표준체와 간편체에 속하지 않는 자

3-5 설명변수

본 연구의 중요한 차별점은 표본코호트 DB에 포함된 개별 코호트의 질병의 진단, 약제의 처방, 검사 및 수술 등 의료서비스 이용 정보를 광범위하게 고려하고, 데이터에 기초한 변수 선택 전략을 채택한 것이다. 설명변수의 영역은 크게 ① 연령정보, ② 의료기관 이용 정보, ③ 과거병력 및 건강문제로 구분해 설명변수 후보군을 설정하였다. 먼저, 연령정보는 자격 및 출생 테이블을 통해 추출하여 5세 단위로 연령집단을 구분해 범주화하였다. 의료이용 규모는 입원 및 내원일수, 의료비 지출 규모를 진료 DB 상에서 추출해 정의하였으며 분포를 고려해 로그 변환해 연속형 변수로 정의하였다.

과거병력에 대한 정보는 명세서 및 상병 테이블에서 확인할 수 있는 KCD(한국표준질병사인분류, Korean Standard Classification of Diseases) 코드를 기준으로 정의하였다. 관찰기간에 한 번이라도 KCD 코드를 기준으로 식별되는 질환을 진단받은 이력이 있다면 해당 질환에 대한 과거 질병력이 존재하는 것으로 정의하였다.

약제 처방기록은 ATC(Anatomical Therapeutic Chemical) 코드를 기준으로 식별해 정의하였다. ATC 코드는 약물을 해부학적 위치, 치료적 작용, 약리학적 특성에 따라 체계적으로 분류하는 WHO 관리 체계이다. 마지막으로 처치 및 수술, 검사와 같은 의료적 행위 정보는 의료서비스에 대한 비용 및 진료 행위를 코드화한 건강보험심사평가원(HIRA)의 수가 코드를 기준으로 정의하였다. 이상의 과거 질병의 진단, 약제와 처치 및 검사와 관련한 설명변수는 이진변수화 하였다.

이처럼 KCD 코드, ATC 코드, 수가코드 보유 유무에 따라 설명변수를 정의하는 경우 변수 규모가 폭증하는 문제를 유발 할 수 있다. 이는 고차원 데이터에 따른 차원 문제, 다중공선성, 그리고 모델의 과적합을 유발할 수 있고 임의적으로 변수를 선택하는 경우 선택편의가 발생할 수 있다. 따라서 본 연구에서는 통계적, 머신러닝적 관점에서 체계적인 변수선택 과정을 시행하였다.

먼저, 발생빈도가 지나치게 낮은 설명변수는 분석에 앞서 제거하였으며, 결과변수와의 단변량 분석을 통해 Information Value(이하 IV)를 산출하여 예측모형 개발에 투입할 설명변수의 후보군을 선정하였다. IV는 범주화된 설명변수가 결과변수를 얼마나 효과적으로 구분하는 지를 수치화한 지표로 개인신용평가 모형 등 다양한 분야에서 변수 선별에 널리 활용된다. 이 값은 각 범주 내 질병 발생자와 비발생자의 비율 차이에 기반하여 계산되며, 두 집단의 분포가 유사할수록 IV는 낮고, 특정 범주에 질병 발생자가 집중될수록 높게 나타난다. 일반적으로 IV 값이 0.02 이상이면 예측 변수로서 유의미한 구분력을 갖는 것으로 간주되며[16]-[18], 본 연구에서도 이를 기준으로 설명변수를 선별하였다. 최종적으로 71개의 변수가 설명변수로 압축되었으며 사전적 변수 선정과정을 통해 도출한 변수 추출과 단변량 분석과 IV를 통해 선정한 설명변수 규모는 표 3과 같다.

이상의 과정을 통해 선정한 설명변수 후보군은 최종적으로 Lasso 로지스틱 회귀를 통해 최종적인 변수선택과 예측모형을 도출하였다. 예측모형을 통해 선택된 변수는 4장 분석결과의 그림 2를 통해 제시한다.

표 3. 변수 선택 단계별 변수 수
Table 3. Number of variables selected during feature selection

Type of EMR data	Step 1	Step 2
Past disease	550	89
Drug prescription	221	115
Tests & interventions	66	32

3-6 예측모형

본 연구에서는 대표적인 선형모형 기반의 머신러닝 방법론인 Lasso 로지스틱 회귀를 활용해 설명변수 선택과 예측을 시도하였다. Tibshrian이 제안한 Lasso는 설명변수 증가에 따른 차원 문제와 변수 임의 선택으로 발생하는 모형 설정의 편의를 완화하는 대표적인 방법론이다[19]. 광범위한 설문조사 자료와 같이 많은 정보들이 수집된 고차원 자료에서 어떤 정보가 유의미한 정보인지를 파악할 때 유용성이 높다[20]. 특히 어떤 설명변수가 결과변수에 대해 예측력을 보유한 변수인지 식별하는 보건의료, 사회과학 분야에서 주로 활용된다. 로지스틱 회귀모델에 벌점을 부과한 Lasso 모델은 아래 식 (1)의 목적함수를 최소화하는 β_0와 β를 추정한다.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i}^{N} l(y_i, \beta_0 + X_i\beta) + \lambda \sum_{j}^{k} |\beta_j| \quad (1)$$

여기서 $l(y_i, \beta_0 + X_i\beta)$은 로지스틱 회귀모델의 로그우도함수를 의미하며, $\lambda \sum_{j}^{k} |\beta_j|$은 k개의 변수가 있을 때 적용되는 L1 벌점항으로 일부 회귀계수의 추정값을 0으로 수축시켜 예측에 중요하지 않은 변수를 제거하는 역할을 한다. 다시 말해, Lasso 로지스틱 회귀모델은 많은 후보군에서 중요한 변수를 선택하고 예측모형의 과적합을 방지하고 예측 성능을 개선하는 과정을 동시에 수행할 수 있는 머신러닝 기법이다. 이러한 장점으로 질환의 발생을 예측하는 과정에서 발병자 수가 적은 사건을 예측해야 하는 경우 일반적인 회귀모델 보다 예측력이 뛰어나며 모델의 일반화가 용이하다[21].

여기서 벌점 모수(penalty parameter) λ를 최적화하는 과정이 모델의 훈련 과정이다. 벌점 모수(λ) 값의 크기에 따라 변수축소 정도가 달라지므로 교차검증(K-fold 교차검증)을 통해 예측성능과 모델의 변수축소 사이의 균형을 최적화하는 방식을 취한다. 이를 위해 전체 데이터를 모형 학습을 위한 학습데이터와 모델의 성능과 변수축소의 최적화를 위한 검증데이터로 8:2의 비율로 분할하였다. 이 과정에서 결과변수의

분포를 훈련용 표본과 검증용 표본에서 균형 있게 유지하기 위해 표본층화추출 방식을 적용하였다. 단순 랜덤 샘플링을 적용할 경우, 질환 발생 여부의 비율이 훈련용 표본과 검증용 표본 간에 불균형하게 나타날 가능성이 있다. 모델 학습에는 훈련용 표본만을 활용해 수행하며 이 과정에서 10-Folds 교차검증을 적용하였다. 이는 훈련용 표본을 10개의 Fold로 나눈 뒤 9개의 Fold로 모델을 학습하고 1개의 Fold로 모델을 평가하는 과정에서 Fold를 교차해 반복적으로 수행하는 방법이다. 이 과정에서 최적의 성능을 산출하는 벌점항의 매개변수 λ의 최적값을 도출하였다. 본 연구에서는, 각 예측요인이 질병 발생의 위험도를 높이는 방향으로 작용하도록, 모든 계수는 음수가 되지 않도록(non-negative) 제약을 추가했다.

Ⅳ. 분석 결과

4-1 예측요인 선정 결과

본 연구는 예측대상 질환과 성별을 구분하여 총 12개의 개별 예측모형을 도출하였다. 각 모형에서는 Lasso 로지스틱 회귀를 적용하여 약물 처방, 검사, 처치, 의료 이용 패턴 등 방대한 설명변수 중 예측 성능에 유의하게 기여하는 요인을 통계적으로 선별하였다. 표 4는 이러한 분석을 통해 도출된 상위 10개의 핵심 예측 변수를 요약한 것으로 성별에 따라 변수 구성의 차이가 나타난다. 주요 만성질환 이력(고혈압, 당뇨, 고지혈증 등)과 의료 이용 행태(통원횟수, 응급실 방문 이력 등)가 포함된다. 이는 모형의 임상적 해석 가능성과 보험 리스크 평가의 실제 활용도를 높이는 데 기여한다.

그림 2는 이러한 변수선택 결과를 '선택비율(Selection Ratio)'로 나타낸 것이다. 세로축에는 최종적으로 선택된 변수들의 목록을, 가로축에는 변수별 선택비율(0~1)을 표시하였다. 선택비율이 1이면 해당 변수가 모든 모형에서 선택되었음을, 0이면 어느 모형에서도 선택되지 않은 경우를 의미한다. 여기서 강조하고 싶은 점은, 변수 선택 과정이 의료적 판단이나 임상적 기준이 아니라 데이터 분석 관점에서 이루어졌다는 것이다. 즉, 예측성능 기여도가 높은 변수가 우선적으로 포함되었다는 점을 밝혀둔다.

주요 결과를 살펴보면, 먼저 고혈압·당뇨·고지혈증과 같은 만성질환 진단 이력이 0.8~1 사이의 높은 선택비율을 보였다. 이는 성별 차이가 존재하긴 하지만, 암·뇌·심장질환 발생 예측에서 이러한 만성질환 변수가 중요한 역할을 한다는 점을 의미한다. 실제로 해당 질환들은 일반적으로 보험계약 시 고지 대상으로 포함되는 질환 이력이기도 하다.

또한 약물처방 관련 변수(향진균제, 항혈전제, 비타민류 등)들도 0.5 이상의 비교적 높은 선택비율을 기록하였다. 이는 특정 약물 복용이 개인의 기저 건강상태나 위험도를 일부 대리함으로써, 향후 중증 질환 발생을 예측하는 유의미한 지표가 될 수 있음을 시사한다.

표 4. 다빈도 관측 변수 목록
Table 4. List of frequently observed variables

Variable	Selected outcomes (Female)	Selected outcomes (Male)
[Diagnosis] Hypertension	y01, y03, y04, y05, y06	y01, y02, y03, y04, y05, y06
[Prescription] Antifungal	y01, y02, y03, y05, y06	y02, y03, y05
[Usage] Outpatient Days	y01, y02, y03, y05, y06	y02, y03, y05
[Examination] Tumor image test	y03, y04, y05, y06	y03, y04, y05, y06
[Diagnosis] Hyperlipidemia	y02, y03, y05, y06	y03, y05, y06
[Usage] Emergency Care	y03, y04, y05, y06	y03, y04, y05
[Diagnosis] Arrhythmia	y03, y04, y05, y06	y03, y04, y05
[Examination] Opthalmology	y01, y02, y04, y05	y01, y02, y04
[Diagnosis] Diabetes	y01, y03, y06	y01, y03, y04, y05, y06
[Prescription] Vitamin	y02, y03, y05	y01, y02, y03, y05, y06

y01 :General Cancer, y02 : Carcinoma in situ
y03 : Cerebrovascular, y04 : Stoke
y05 : Ischemic Heart Disease
y06 : Acute Myocardi-al Infarction

아울러 의료이용 규모와 응급서비스 경험(통원횟수, 응급실 방문 이력 등) 역시 다양한 모형에서 선택비율이 0.5 수준으로 나타났다. 이는 질환 진단·약물 복용 이력만이 아니라, 의료서비스 접근 행태도 중장기 질환 발생에 대한 예측력을 가질 수 있음을 의미한다. 이때 세부 변수의 중요도나 선택비율은 여성과 남성 집단 간에 다소 차이를 보이기도 한다.

흥미롭게도, 어지러움증·두통·경증 호흡기 질환과 같이 임상적으로 가벼운 증상으로 분류되는 진단도 약 0.5 전후의 선택비율을 나타냈다. 임상적으로 큰 위험요인이 아니더라도, 타 질환의 증상으로 동반되어 나타날 경우 중장기 질환 발생과 일정한 상관관계를 보일 수 있음을 시사한다. 다만 본 연구는 데이터 분석 관점에서 예측력을 최우선시하였으므로, 해당 변수들에 대한 임상적 인과관계는 추가 연구가 필요하다.

마지막으로, 몇몇 변수들은 여성 모형에서만 선택되거나, 반대로 남성 모형에서만 두드러진 비율을 보여 성별 간 선택되 변수의 차이가 존재한다. 질환과 성별에 따라 선택된 변수의 개수를 함께 고려하면 예측 대상에 따라 선택된 변수의 규모가 상이함을 확인할 수 있다. 가령, 에스트로겐의 복용은 여성에서만 유의한 요인으로 식별된다.

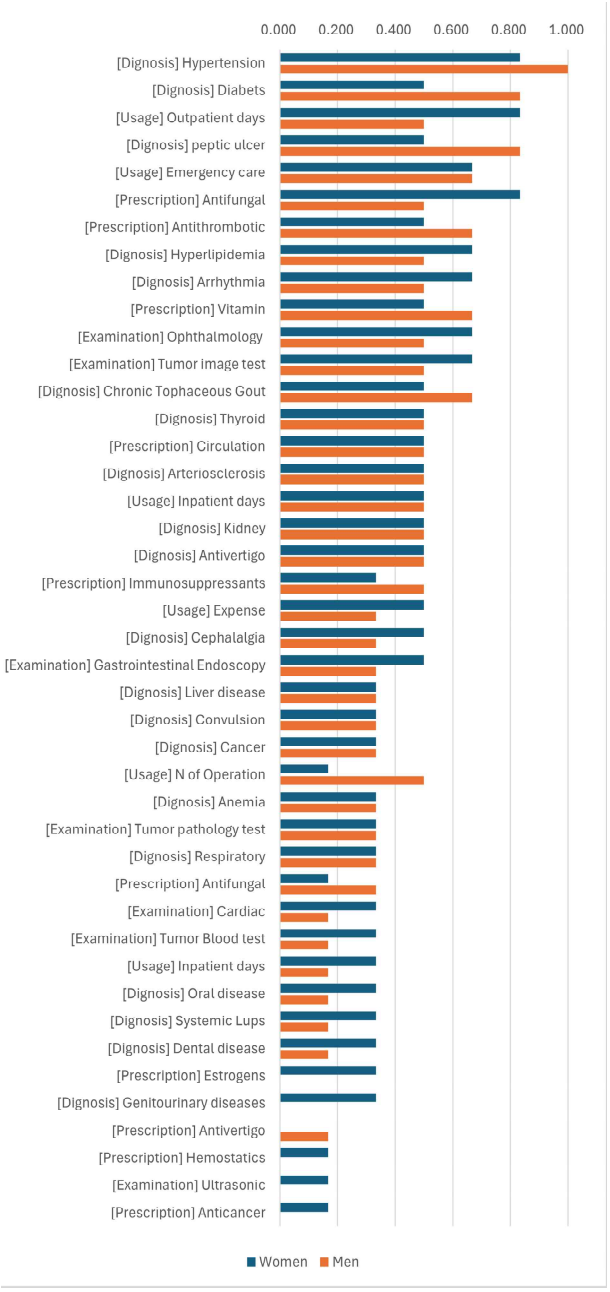

그림 2. 변수별 선택 비율
Fig. 2. Selection ratio by variable

이와 같은 결과는 기존 계약 전 알릴 의무 항목이나 건강검진 정보만으로는 충분히 포착하기 어려웠던 특성들을 위험과 보험료 세분화에 반영할 여지가 있음을 보여준다. 예를 들어, 고지항목이나 건강검진 항목은 고정적이고 보험상품이 보장하는 질환이나 집단에 특화된 요인과 특성을 발굴에 제한적이다. 따라서 위험수준에 따른 수직적 세분화의 성격을 가지게 된다. 반면에 의료이용 내용 정보는 광범위하게 건강과 유관한 정보를 활용함으로써 보장 질환과 가입대상에 대해 차별적인 예측요인과 특성을 고려할 수 있고 이는 수평적 세분화가 가능함을 의미한다. 이러한 수평적 세분화는 동일한 위험수준을 가진 대상이 담보하는 위험과 성별과 같은 특성 조합에 따라 다양한 위험 수준으로 평가받을 수 있음을 의미한다. 이는 보험위험의 불확실성을 관리하여 보험료 부담의 공정성과 형평성을 담보하는 방식으로 고령자나 유병자 대상의 시장 확대가 가능할 수 있음을 의미한다.

4-2 예측모형 성능

선택된 변수에 기반한 예측모형을 구축한 후 검증용 표본을 활용하여 모형의 성능을 평가하였다. 각 모형에 대해 임계값(Threshold)를 적용해 추적기간 내 질환발생을 양성(Positive)으로, 미발생을 음성(Negative)로 분류한 뒤 표 5에 제시한 성능지표를 산출하였다. 본 연구에서는 성능지표로는 AUC(Area Under the ROC Curve), 민감도(Senstivity), 특이도(Specificity), F1 스코어를 주된 평가척도로 삼았다.

표 5의 성능지표 결과를 살펴보면, 모형의 전반적인 예측성능을 평가할 수 있는 AUC 값은 0.7~0.8 수준에서 형성되어 양호한 성능으로 평가된다. AUC는 모형이 질환의 발생과 미발생을 얼마나 잘 분류하는 가를 평가하는 것으로 가능한 모든 임계값에서의 예측력을 전반적으로 보여준다. 반면에 여성 유사암 모형의 경우 AUC 값이 0.6 수준으로 상대적으로 예측력이 떨어지는 것으로 나타났다. 그림 3은 성별 일반암, 뇌혈관질환, 허혈성심장질환 예측모형의 ROC 곡선을 시각화한 것으로, 각 모형의 민감도와 특이도의 균형 수준을 직관적으로 확인할 수 있다.

실제 질환 발생자를 '발생자'로 분류하는 지표인 민감도는 대다수 모형에서 0.8 이상의 수치를 보이고 있다. 즉, 질환이 발생한 대상자를 놓치는 거짓음성 경우가 낮은 편으로 위험군 식별 측면에서 유의미한 성능을 보이고 있다. 반면 질환 미발생자를 '미발생 자'로 분류하는 거짓양성에 대한 특이도는 0.5~0.75 수준으로 민감도 대비 상대적으로 낮게 나타난다. 이는 거짓양성(False Positive)이 발생할 여지가 있어, 미발생자를 발생자로 분류하는 경향이 존재할 수 있음을 의미한다.

마지막으로, 정밀도는 예측된 발생 사례 중 실제로 발생한 사례의 비율을 나타내는 지표로, 예측 결과의 정확성을 평가하는 데 활용된다. 본 연구에서 도출된 정밀도는 전반적으로 낮은 수준(0.007~0.114)을 보였으며, 일부 질환에서는 0.01 내외에 불과하였다. 이러한 결과는 암, 뇌, 심혈관질환과 같은 희귀 질환의 예측에서 거짓양성이 다수 발생했음을 의미한다. 민감도를 높이기 위해 실제 발생하지 않은 사례도 발생으로 예측하게 되면서, 결과적으로 정밀도가 낮아진 것이다.

예측모형의 성능을 종합적으로 평가하면, AUC와 민감도는 높게 나타났으나 정밀도와 특이도는 낮아 '민감도 중심의 예측모형' 성격을 보인다고 할 수 있다. 이는 질환 발생 가능성이 높은 고위험군을 우선적으로 식별하는 상황에서는 긍정

표 5. 예측모형 성능 지표
Table 5. Predictive model metrics

Disease	Gender	No. of Selected vars	Threshold	AUC	Sensitivity	Specificity	Precision
General cancer	F	13	0.016	0.741	0.876	0.491	0.029
	M	11	0.020	0.844	0.844	0.713	0.060
Carcinoma in Situ	F	15	0.007	0.682	0.881	0.397	0.011
	M	15	0.005	0.771	0.789	0.633	0.010
Cerebrovascular	F	25	0.045	0.844	0.802	0.748	0.114
	M	24	0.028	0.836	0.835	0.682	0.083
Stroke	F	17	0.003	0.792	0.810	0.655	0.007
	M	14	0.003	0.774	0.833	0.599	0.007
Ischemic heart disease	F	23	0.031	0.822	0.803	0.714	0.078
	M	21	0.031	0.799	0.803	0.656	0.074
Acute myocardial infarction	F	19	0.002	0.860	0.848	0.741	0.008
	M	16	0.005	0.811	0.849	0.660	0.011

적으로 해석될 수 있다. 예를 들어 보험산업 관점에서는 질환 발생 가능성이 높은 가입자를 조기에 선별함으로써 잠재적 리스크 노출을 최소화할 수 있기 때문이다. 다만 이러한 모형은 위험을 과도하게 추정하여 보험계약 접근성을 저해할 가능성도 있다. 낮은 정밀도의 문제를 완화하기 위해 재심사 또는 보완 평가 절차를 도입하는 등의 전략이 필요하다. 결국, 민감도와 정밀도 간의 균형을 어떻게 설정하느냐에 따라 모형의 활용 방향과 실무적 용도가 결정될 수 있다.

4-3 기준시점과 추적기간에 따른 예측모형 성능검증

본 연구의 예측모형은 2015년 1월 1일을 기준시점으로 정의하고 과거 5년을 관찰기간, 이후 5년을 추적기간으로 정의한 학습데이터를 기반으로 한다. 예측모형이 특정 기준시점과 기간에 과도하게 최적화되는 경우 기준시점과 추적기간의 변화화에 일관된 성능을 유지하기 어렵다. 이는 실무적 관점에서 중요한 이슈로 동일한 시점과 기간에 기반한 교차검증 이외에 시점과 기간을 변경해 예측모형의 성능을 평가하는 절차가 요구된다. 이에 기준시점 변경에 따른 검증과 추적기간 변경에 따른 검증을 구분해 수행하였다. 관찰 기간의 경우 앞서 Ⅲ장에서 언급한 바와 같이 보험의 계약전 알릴사항의 고지기간을 고려하여 기준시점 이전 5년으로 고정하여 일관되게 적용하였다.

기준시점 변경에 따른 성능 검증은 전체 코호트를 대상으로 기준시점을 2010년, 2013년, 2015년으로 각각 설정해 앞서 구현한 예측모형에 대응하는 입력변수 구성하고 이를 예측모형에 입력해 산출된 예측값과 실제 데이터 상의 관측값을 비교해 AUC 값을 산출하였다. 이를 통해 만약 기준시점 변경으로 관찰기간이 과거로 갈수록 의료이용 패턴이나 청구 및 치료 관행의 변화 등과 같은 시간적 불안정과 예측모형의 변수 선택의 시간적 비일관성의 존재로 발생할 수 있는 기준시점별 예측성능의 변화를 검토하였다.

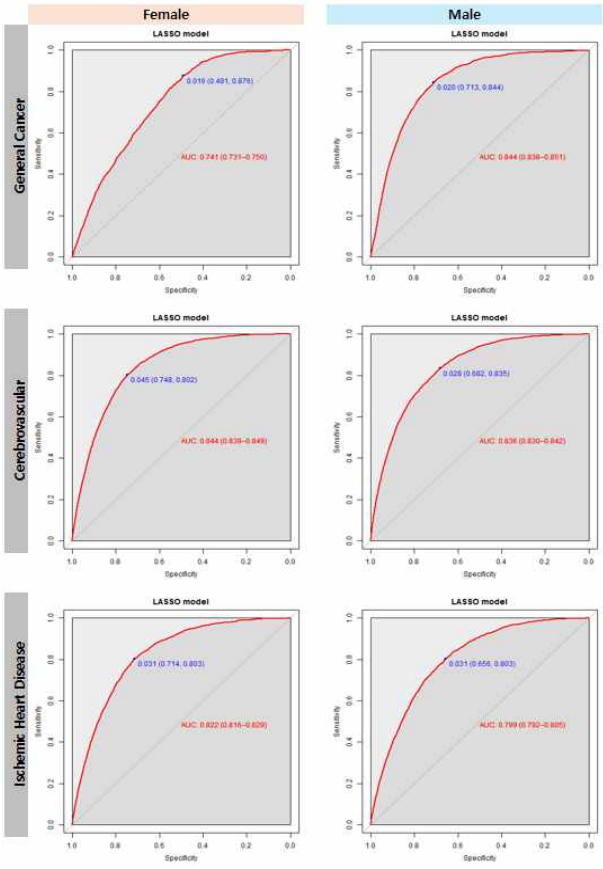

그림 3. 예측모형 성능 평가를 위한 ROC 곡선
Fig. 3. ROC curve for evaluating model performance

추적기간에 따른 성능 검증은 기준시점을 2015년으로 고정한 상태에서 추적기간을 1년에서 4년까지 다양하게 설정하였다. 전체 표본 코호트를 대상으로 예측모형에 대응하는 입력변수를 구성하고 이를 예측모형의 입력하여 예측값을 산출하고 추적기간별 실제 데이터 상의 관측값 비교해 예측 성능을 비교하였다. 추적기간이 짧을수록 예측대상 질환의 발생률

은 낮아지고, 그에 따라 예측대상의 불균형이 심해 질 수 있다. 또한 의료이용 이력은 단기 발생 위험 예측에 유효해 추적기간이 길어질수록 예측력이 감소할 수 있다. 이 점을 고려하여 추적기간에 대한 검증을 통해 예측모형의 장-단기 예측에 대한 안정성을 평가하고자 하였다.

표 6은 기준시점의 변경에 따른 예측모형의 성능을 비교한 결과를 보여 준다. 기준시점에 따라 AUC값을 고려할 때 기준시점에 따른 모형 성능의 변화는 크지 않다. 각 성별과 예측대상 질환에 대해서도 대해 예측성능의 큰 저하는 관찰되지 않는다. 이는 예측성능 관점에서 본 연구의 예측모형의 설명변수 선택이 기준시점에 대한 일관성을 보유하고 있으며, 과거 자료의 시간적 불안정성에 따른 영향이 크지 않다는 점을 시사한다.

표 7은 기준시점을 2015년 1월 1일로 고정한 상태에서, 추적관찰 기간을 연차별(1차년~4차년)로 구분하여 예측모형의 성능을 비교한 결과를 보여준다. 추적기간에 따른 AUC 값 또한 모형의 성능 변화가 크지 않은 점을 확인 할 수 있다. 다만 예측성능은 추적기간이 길어짐에 따라 감소하고 있는데 이는 설명변수의 예측력이 장기적으로 감소할 수 있음을 시사한다. 따라서 예측모형이 5년 이내의 추적기간에 대한 예측에서 모두 성능을 유지하고 있지만 5년 이상의 장기 예측시 주의가 필요하다.

표 6. 기준시점에 따른 데이터별 성능지표 검증
Table 6. Time-point validation results by baseline

Category		Time-point based validation		
Disease	Gender	2015	2013	2010
General cancer	F	0.7407	0.7520	0.7604
	M	0.8444	0.8528	0.8565
Carcinoma in Situ	F	0.6817	0.7027	0.7038
	M	0.7712	0.7543	0.7484
Cerebrovascular	F	0.8439	0.8519	0.8603
	M	0.8357	0.8459	0.8522
Stroke	F	0.7919	0.7907	0.8023
	M	0.7744	0.7799	0.7770
Ischemic heart disease	F	0.8222	0.8305	0.8341
	M	0.7987	0.8121	0.8167
Acute myocardial infarction	F	0.8600	0.8800	0.8432
	M	0.8106	0.8103	0.8093

표 7. 연차별 예측 효과 검증 결과
Table 7. Validation of prediction performance by year

Category		Follow-up Year Validation			
Disease	Gender	1st year	2nd year	3rd year	4th year
General cancer	F	0.756	0.736	0.737	0.717
	M	0.859	0.839	0.837	0.833
Carcinoma in Situ	F	0.732	0.705	0.690	0.673
	M	0.846	0.774	0.755	0.733
Cerebrovascular	F	0.852	0.860	0.849	0.844
	M	0.853	0.848	0.845	0.832
Stroke	F	0.777	0.771	0.784	0.811
	M	0.802	0.761	0.779	0.748
Ischemic heart disease	F	0.837	0.843	0.835	0.834
	M	0.829	0.822	0.816	0.805
Acute myocardial infarction	F	0.862	0.879	0.874	0.865
	M	0.830	0.819	0.809	0.787

4-4 상대위험도 산출과 검토

앞서 도출한 예측모형을 기반으로 성별과 연령집단에 따른 예측발생률을 산출하고 이를 활용해 보험 가입유형에 따른 상대위험도를 계산하였다. 상대위험도는 특정 집단(예: 간편체 또는 거절체)의 '평균 발생률'을 표준체 대비 비율로 나타낸 것이며, 각 집단 간 질병 발생 위험의 상대적 크기를 직관적으로 파악할 수 있는 지표이다. 상대위험도 분석은 보험료 세분화와 밀접한 의미를 지닌다. 즉, 각 집단별로 질환 발생 위험에 차이가 있다면, 그 차이를 보험료 혹은 담보 설계에 반영함으로써 공정한 보험료 산정과 보험위험 감소가 가능해진다.

표 8과 표 9는 상대위험도 산출을 위해 예측모형으로 산출한 예측 발생률을 보여준다. 예측 발생률은 각 예측질환과 성별에 따른 예측모형의 추정값을 사용하여 개인별 질환 예측 발생률을 구한 다음, 성별 및 연령군별 평균을 계산하여 산출하였다. 또한, 데이터 상에서 실제 관측된 발생률도 함께 제시하였다. 표 8과 표 9를 통해 예측 발생률과 관측 발생률을 살펴보면 각 질환과 인구집단에 대해 예측 발생률과 관측된 발생률 차이가 크지 않음을 확인할 수 있다. 이는 개인별 예측 오차가 집단 수준에서는 상쇄되어 집단에 대한 예측 타당성이 확보됨을 시사한다. 이는 개인 단위 예측에서의 오차에도 불구하고, 보험 정책 수립이나 집단 위험 평가 등 실무적 활용에는 예측모형이 충분한 신뢰성을 가질 수 있음을 시사한다.

또한 예측 발생률의 패턴을 보면 연령집단이 증가함에 따라 함께 증가하는 전형적인 형태가 나타나고 있다. 45세~49세 연령집단을 시작으로 대다수 질환에서 발생률이 가파르게 상승하는 패턴이 관측된다. 이러한 패턴은 관측된 발생률에서도 동일하게 나타나고 있다. 또한 예측과 관측 발생률 모두에서 심장질환에서 남성이 뚜렷하게 높은 발생률을 보이고 있으나 암이나 뇌질환에서는 여성 발생률이 오히려 높게 나타나 질환과 성별에 따른 차이가 존재함을 확인할 수 있다.

표 10과 표 11은 질환별로 표준체 대비 간편체 및 거절체의 상대위험도를 제시한 것이다. 상대위험도는 성별 및 연령군별로, 간편고지 대상자 또는 거절체 대상자의 질환 예측 발생률을 표준체의 평균 예측 발생률로 나눈 값으로 정의하였다. 이러한 방식으로 산출된 개인별 상대위험도는 성별·연령군·가입유형별로 평균하여 비교하였다.

표 8. 연령별 질병 발생 예측(여성)
Table 8. Age-specific disease incidence prediction (female)

Group	General cancer		Carcinoma in Situ		Cerebrovascular		Stroke		Ischemic heart disease		Acute myocardial infarction	
	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs
20–29	0.0030	0.0031	0.0036	0.0037	0.0049	0.0051	0.0007	0.0007	0.0054	0.0054	0.0002	0.0002
30–39	0.0104	0.0105	0.0079	0.0082	0.0090	0.0090	0.0009	0.0010	0.0079	0.0082	0.0003	0.0003
40–44	0.0183	0.0185	0.0088	0.0093	0.0173	0.0172	0.0015	0.0016	0.0157	0.0158	0.0008	0.0008
45–49	0.0236	0.0232	0.0103	0.0101	0.0307	0.0311	0.0030	0.0030	0.0249	0.0249	0.0011	0.0010
50–54	0.0237	0.0240	0.0100	0.0102	0.0474	0.0472	0.0037	0.0036	0.0408	0.0406	0.0020	0.0021
55–59	0.0267	0.0269	0.0100	0.0102	0.0672	0.0677	0.0043	0.0045	0.0561	0.0558	0.0023	0.0022
60–64	0.0333	0.0326	0.0104	0.0104	0.0934	0.0936	0.0049	0.0050	0.0759	0.0759	0.0038	0.0044
65–69	0.0406	0.0394	0.0119	0.0114	0.1233	0.1226	0.0066	0.0069	0.0964	0.0948	0.0064	0.0067
70+	0.0471	0.0471	0.0106	0.0105	0.1743	0.1743	0.0123	0.0122	0.0969	0.0975	0.0134	0.0133

표 9. 연령별 질병 발생 예측(남성)
Table 9. Age-specific disease incidence prediction (male)

Group	General Cancer		Carcinoma in Situ		Cerebrovascular		Stroke		Ischemic heart disease		Acute myocardial infarction	
	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs
20–29	0.0016	0.0016	0.0014	0.0013	0.0043	0.0042	0.0007	0.0008	0.0068	0.0070	0.0004	0.0004
30–39	0.0047	0.0048	0.0035	0.0034	0.0104	0.0109	0.0016	0.0016	0.0160	0.0161	0.0016	0.0015
40–44	0.0088	0.0090	0.0039	0.0039	0.0188	0.0189	0.0027	0.0026	0.0279	0.0276	0.0036	0.0036
45–49	0.0150	0.0151	0.0047	0.0044	0.0299	0.0299	0.0039	0.0041	0.0399	0.0389	0.0054	0.0054
50–54	0.0238	0.0243	0.0061	0.0063	0.0488	0.0479	0.0049	0.0050	0.0547	0.0547	0.0078	0.0079
55–59	0.0402	0.0408	0.0078	0.0082	0.0655	0.0660	0.0062	0.0064	0.0706	0.0707	0.0097	0.0098
60–64	0.0607	0.0610	0.0104	0.0103	0.0920	0.0918	0.0072	0.0074	0.0905	0.0892	0.0125	0.0121
65–69	0.0864	0.0875	0.0123	0.0122	0.1217	0.1205	0.0077	0.0082	0.1070	0.1074	0.0120	0.0123
70+	0.1097	0.1097	0.0137	0.0136	0.1696	0.1700	0.0151	0.0149	0.1049	0.1058	0.0178	0.0182

표 10. 표준체 대비 간편체 상대위험도
Table 10. Relative risk (simplified vs. standard) by age and gender

	General cancer		Carcinoma in Situ		Cerebrovascular		Stroke		Ischemic heart disease		Acute myocardial infarction	
Age group	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
20–29	1.188	1.064	1.227	1.224	1.306	1.216	1.205	1.153	1.322	1.252	1.249	1.141
30–39	1.127	1.094	1.208	1.217	1.241	1.268	1.124	1.207	1.302	1.330	1.225	1.189
40–44	1.123	1.098	1.220	1.252	1.333	1.317	1.191	1.245	1.422	1.418	1.243	1.209
45–49	1.123	1.116	1.229	1.277	1.400	1.364	1.235	1.281	1.527	1.491	1.309	1.244
50–54	1.126	1.136	1.230	1.311	1.452	1.427	1.276	1.335	1.607	1.581	1.376	1.291
55–59	1.131	1.152	1.229	1.338	1.494	1.488	1.315	1.380	1.678	1.676	1.450	1.338
60–64	1.144	1.167	1.234	1.356	1.526	1.506	1.355	1.416	1.721	1.704	1.524	1.374
65–69	1.154	1.167	1.236	1.366	1.548	1.521	1.387	1.453	1.776	1.723	1.588	1.407
70+	1.173	1.188	1.247	1.379	1.631	1.565	1.503	1.554	1.889	1.825	1.779	1.479

표 10의 결과에 따르면, 간편체는 전 연령·성별 집단에서 표준체 대비 상대위험도가 1을 상회하며, 상대적으로 질환 발생률이 높게 나타난다. 또한 표 11에서 거절체는 모든 질환에서 높은 상대위험도를 기록하여, 보험사의 기존 고지기반 분류와 예측모형 기반 분류 간의 일관성을 확인할 수 있다.

간편체는 표준 고지항목 일부가 면제된 상태에서 가입이 허용된 집단으로, 실제로 과거 질환 이력이나 의료이용 경향이 상대적으로 높을 가능성이 있다. 따라서 이들 집단에서 높은 상대위험도가 산출된 결과는 보험사의 분류체계와 데이터 기반 예측 결과 간의 정합성을 시사한다.

표 11. 표준체 대비 거절체 상대위험도
Table 11. Relative risk (declined vs. standard) by age and gender

	General cancer		Carcinoma in Situ		Cerebrovascular		Stroke		Ischemic heart disease		Acute myocardial infarction	
Age group	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
20-29	1.278	1.168	1.404	1.428	1.680	1.586	1.540	1.396	1.873	1.742	1.629	1.426
30-39	1.267	1.256	1.451	1.360	1.672	1.691	1.484	1.504	1.879	1.862	1.414	1.513
40-44	1.227	1.228	1.453	1.416	1.664	1.760	1.478	1.552	1.916	1.910	1.483	1.457
45-49	1.231	1.250	1.440	1.470	1.756	1.843	1.545	1.603	1.981	2.063	1.538	1.556
50-54	1.215	1.268	1.403	1.501	1.787	1.889	1.558	1.653	2.034	2.142	1.597	1.621
55-59	1.214	1.268	1.399	1.518	1.817	1.917	1.583	1.674	2.064	2.207	1.643	1.672
60-64	1.222	1.268	1.377	1.517	1.836	1.893	1.614	1.699	2.161	2.185	1.769	1.692
65-69	1.232	1.270	1.402	1.541	1.875	1.901	1.659	1.721	2.281	2.231	1.893	1.737
70+	1.245	1.275	1.393	1.514	1.991	1.901	1.806	1.814	2.437	2.325	2.223	1.846

V. 결 론

본 연구는 국민건강보험 표본코호트 DB에 축적된 세부 의료이용 내역을 활용하여 암·뇌·심장질환 등 대표적인 3대 만성질환의 발생을 예측하는 머신러닝 모형을 구현하였다. 특히 데이터 기반 변수선택 전략을 통해 질환 발생의 주요 예측요인을 식별하고, 집단별 상대위험도를 분석함으로써 보험산업 관점에서 예측모형의 의미를 논의하였다. 연구 결과와 시사점을 정리하면 다음과 같다.

먼저, 예측 대상으로 삼은 질환(암·뇌·심장)에 따라 중요하게 선택된 설명변수가 달랐다. 이는 기존의 계약 전 알릴 의무 항목이나 건강검진 정보만으로는 놓칠 수 있었던 특성을 의료이용 내역을 통해 포착할 수 있음을 시사한다. 특히 질환이나 성별에 따라 위험요인이 다르게 작동한다는 점은, 동일 위험 수준을 지닌 가입자라도 보장 대상 질환·성별 등에 따라 상세한 세분화가 가능하다는 의미다. 결과적으로, 조건에 따라 개인이 다른 위험 수준으로 평가받을 수 있게 되어, 보험료 세분화의 폭이 한층 확대될 여지를 보여준다.

예측모형이 질환 발생을 놓치지 않도록(고위험군 포착) 설계될 경우 민감도는 높아지지만, 저위험군을 고위험군으로 분류하는 거짓양성 가능성도 함께 증가한다. 개인단위 예측을 시도하는 경우 보험료 부담의 형평성에 부정적일 수 있다. 따라서 민감도와 특이도 사이의 상충관계를 고려할 때 목적과 실무적 용도에 따라 절충점을 찾는 것이 요구된다.

보험가입 유형에 따른 상대위험도의 체계적 차이는 기존 분류체계의 타당성을 뒷받침하며, 예측모형이 이를 정밀하게 재현할 수 있음을 시사한다. 이는 기존 분류체계와 크게 상충되지 않음을 보여주는 동시에, 의료이용 정보와 예측모형을 활용하면 이미 고위험군으로 분류된 가입자들 내에서도 상대적 위험도를 더욱 세밀하게 구분할 수 있음을 의미한다. 따라서 새로운 위험집단을 발굴하기보다는, 고위험군 내부의 위험도 차이를 반영하여 차등구조를 세분화하고, 보험료 산정 기준에도 이를 반영하는 것이 실질적 보험 설계에 유효할 수 있다.

종합하면 보건의료 빅데이터와 머신러닝 기술을 접목한 예측모형은 소비자에게는 맞춤형 보장 기회를, 보험사에는 효율적인 위험 관리 전략의 기반을 제공할 것으로 기대된다. 나아가 인슈어테크 기술의 확산은 개인정보 보호, 임상적 타당성 확보, 소비자 수용성 등 복합적 쟁점을 수반하므로, 관련 제도와 윤리적 기준에 대한 체계적인 검토가 병행되어야 한다.

참고문헌

[1] S. H. Jeong, S. C. Hong, J. Y. Lee, S. H. Hwang, and H. J. Moon, The Role and Challenges of National and Private Health Insurance, Korea Insurance Research Institute, Seoul, Technical Report No. 2022-01, 2022.

[2] S. J. Kim, T. J. Lim, and Y. M. Kim, The Impact of Population Aging on Health Insurance Demand: A Quantitative Analysis Using a Structural Model, Korea Insurance Research Institute, Seoul, Technical Report No. 22-08, August 2022.

[3] H. J. Shin, "A Study on Domestic and International Trends and Prospects of InsurTech," *The Journal of Trade and Insurance*, Vol. 21, No. 1, pp. 55-68, 2020. https://doi.org/10.22875/jiti.2020.21.1.004

[4] Financial Services Commission (FSC). Digital Insurance Innovation Strategy for the Future [Internet]. Available: https://www.fsc.go.kr/no010101/84163

[5] H. J. Chun and T. K. Leen, "Predicting Relative Risk of Cancer Occurrence and Treatment Based on Simplified Issue Disclosure Period," *Korean Insurance Journal*, Vol. 138, pp. 41-72, April 2024. https://doi.org/10.17342/KIJ.2024.138.2

[6] J. Hippisley-Cox, C. Coupland, and P. Brindle, "Development and Validation of QRISK3 Risk Prediction

Algorithms to Estimate Future Risk of Cardiovascular Disease: Prospective Cohort Study," *BMJ*, Vol. 357, j2099, May 2017. https://doi.org/10.1136/bmj.j2099

[7] G. F. Stark, G. R. Hart, B. J. Nartowt, and J. Deng, "Predicting Breast Cancer Risk Using Personal Health Data and Machine Learning Models," *PLoS ONE*, Vol. 14, No. 12, e0226765, December 2019. https://doi.org/10.1371/journal.pone.0226765

[8] S. C. Hong, S. Y. Lee, S. H. Kim, and S. H. Jun, "Predicting the Future Disease Burden and Medical Cost: Using National Health Insurance Big Data," *The Korean Journal of Economic Studies*, Vol. 71, No. 2, pp. 5-55, 2023. https://doi.org/10.22841/kjes.2023.71.2.001

[9] J. H. Yu, S. H. Kwon, C. M. B. Ho, K. R. Lee, N. S. Kim, C. S. Pyo, and S. J. Park, "Stroke Disease Prediction Based on Deep Learning Using the Elderly Cohort DB," *Journal of Digital Contents Society*, Vol. 21, No. 6, pp. 1191-1200, June 2020. https://doi.org/10.9728/dcs.2020.21.6.1191

[10] T. H. Lim and S. C. Han, "Research on a Diagnostic Model of Deep Learning-Based Pneumonia Using Defense Medical Data," *Journal of Digital Contents Society*, Vol. 22, No. 3, pp. 509-517, March 2021. https://doi.org/10.9728/dcs.2021.22.3.509

[11] D. G. Lee, K. K. Byun, H. D. Lee, and S. H. Shin, "The Prediction of Survival of Breast Cancer Patients Based on Machine Learning Using Health Insurance Claim Data," *Journal of the Korea Industrial Information Systems Society*, Vol. 28, No. 2, pp. 1-9, April 2023. https://doi.org/10.9723/jksiis.2023.28.2.001

[12] P. J. Oh, H. C. Kim, and H. S. Kwon, "A Method for Evaluating and Scoring of Health Status," *The Korean Journal of Applied Statistics*, Vol. 33, No. 3, pp. 239-256, June 2020. https://doi.org/10.5351/KJAS.2020.33.3.239

[13] S. Y. Lee, E. S. Jo, S. H. Jun, and S. C. Hong, "Predicting the Risk of Major Chronic Diseases and Its Application: Using NHIS Big Data," *Korean Insurance Journal*, Vol. 133, pp. 23-48, January 2023. http://dx.doi.org/10.17342/KIJ.2023.133.2

[14] H. S. Kwon, P. J. Oh, M. Y. Kang, and K. S. Woo, "A Study on Insurance Premium Rate Differentiation by Simplified Issue Insurance Product Type Using the Health Level Scoring Model based on National Health Insurance Data," *The Journal of Risk Management*, Vol. 32, No. 4, pp. 99-147, December 2021. http://dx.doi.org/10.21480/tjrm.32.4.202112.004

[15] S. H. Han, J. H. Hong, J. H. Choi, S. H. Kim, B. S. Ryu, M. J. Choi, ... and S. C. Hong, "Predicting the Risk of Cardio-Cerebrovascular Disease and Medical Utilization According to Initial Risk Level," *Korean Insurance Journal*, Vol. 140, pp. 27-58, 2024.

[16] R. Anderson, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford: Oxford University Press, 2007.

[17] N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, 1st ed., Hoboken, NJ: Wiley, 2005.

[18] S. Kyeong and J. Shin, "Two-Stage Credit Scoring Using Bayesian Approach," *Journal of Big Data*, Vol. 9, No. 106, 2022. https://doi.org/10.1186/s40537-022-00665-5

[19] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267-288, January 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[20] H. D. Park, "Factors of Simple Payment Service Use: Application of Machine Learning Prediction Model," *Journal of Digital Contents Society*, Vol. 21, No. 5, pp. 921-929, May 2020. https://doi.org/10.9728/dcs.2020.21.5.921

[21] B. N. Mandal and J. Ma, "l_1 Regularized Multiplicative Iterative Path Algorithm for Non-Negative Generalized Linear Models," *Computational Statistics & Data Analysis*, Vol. 101, pp. 289-299, September 2016. https://doi.org/10.1016/j.csda.2016.03.009

류범상(Beom-Sang Ryu)

2019년 : 연세대학교 산업공학 학사

2019년~2021년: SK 하이닉스
2022년~2023년: 주식회사 웰시콘
2024년~현 재: 주식회사 온택트헬스
※관심분야 : 의료데이터 분석(Medical Data Analysis), 헬스케어 인공지능(Healthcare AI) 등

성지민(Ji-Min Sung)

2002년 : Okayama University (일본) 통계학 석사
2005년 : Okayama University (일본) 통계학 박사

2008년~2012년: 연세대학교 의과대학 연구부, 조교수
2012년~2015년: 차의과학대학교 보건복지대학원 초빙교수
2015년~2021년: 연세대학교 의과대학 Connect-AI 센터 연구부교수
2022년~현 재: 연세대학교 의과대학 뇌심혈관질환연구센터 겸임교수
※관심분야 : 임상시험 통계분석, 인공지능 기반 예측모델링, 헬스케어 데이터 사이언스