The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.3.239
33(3), 239–256

A method for evaluating and scoring of health status

Piljae Oh[a] · Hyeoncheol Kim[b] · Hyuksung Kwon[a,1]

[a]Department of Statistics and Actuarial Science, Soongsil University;
[b]Samjong KPMG Digital Consulting

(Received January 7, 2020; Revised February 15, 2020; Accepted March 23, 2020)

Abstract

Health is an important issue due to increased life expectancy. As a result, the demand for industry and services associated with individual health, health-related programs and services will be facilitated by a method to evaluate and classify the health level of an individual based on various factors. This study suggests a methodology to measure and score an individual health level. A credit scoring model was introduced to implement the categorization of variables, construct a prediction model, and to score individual health level. Cohort DB provided by National Health Insurance Service was used to illustrate overall procedures. It is expected that the suggested model can be utilized in designing and managing health care services as well as other health-related programs.

Keywords: Cohort DB, credit scoring model, health status, logistic regression, scoring method

1. 서론

최근 저출산과 기대수명 증가에 따라 고령인구의 비율이 앞으로 지속적으로 증가할 것으로 예측되고 있다. 이에 따라 건강에 대한 관심은 늘어나고 있으며, 건강과 관련한 다양한 정보들이 여러 매체들을 통해 제공되고 있다. 이는 건강에 대한 패러다임이 질병치료 중심에서 질병 예방 및 건강증진 중심으로 변화되고 있음을 의미한다. 다양한 만성질환의 발병은 개인에게는 재정적 손실 및 지속적인 사회활동을 가능하지 못하게 하는 요인이 되고 더 나아가 사회적으로는 공적보험의 의료비 지출이 증가하게 된다.

따라서, 많은 선진국에서는 헬스케어 관련 시장이 급부상하고 있는 상황이며, 우리 정부에서는 최근들어 국가 암 검진 사업 등 건강 검진을 활성화하여 국민의 건강 증진을 위해 노력하고 있다. 일부 보험회사의 경우에는 보험상품 가입자가 충분한 운동량을 달성하는 경우 인센티브를 제공하는 프로그램을 운영하기도 하고, 의료비를 보장해주는 실손형 민영의료보험의 경우 개인의 의료비 관련 보험금 지급 규모에 따라 보험료를 할인 또는 할증하는 제도를 도입하려는 움직임도 나타나고 있다. 개인의 건강상태를 다양한 요인들을 고려하여 측정할 수 있는 합리적인 지표들이 개발되면, 해당 지표를 바탕으로 개인의 건강 향상을 위한 방안을 제시할 수 있고, 이를 건강과 관련한 상품 및 서비스 등의 마케팅에도 활용할 수 있을 것이다.

국민건강보험공단에서는 2014년부터 연구자들에게 건강보험 DB에서 표본 추출한 약 100만명 가량의 건강상태, 질병의 발병 및 사망, 의료이용 등의 정보를 담은 표본 코호트 DB를 제공하고 있다. 현재는

[1]Corresponding author: Department of Statistics and Actuarial Science, Soongsil University, 369 Sangdo-ro, Dongjak-gu, Seoul 06978, Korea. E-mail: hskwon@ssu.ac.kr

표본 코호트 DB 뿐만 아니라 노인 코호트 DB, 건강검진 코호트 DB 등의 데이터를 제공하여 연구자들이 건강 및 질병에 관한 주제들을 대상으로 다양한 연구를 진행하고 있다. 특히, 표본 코호트 DB의 경우 질병 진단 및 사망에 관한 자료와 함께 건강검진 시 측정된 다양한 건강 관련 정보들을 포함하고 있어 건강상태를 측정하고 분류할 수 있는 모형을 설계하는데 유용하게 활용될 수 있을 것으로 판단된다.

따라서, 본 연구에서는 표본 코호트 DB를 이용하여 건강상태를 평가하고 분류할 수 있는 모형을 설계하는 방안을 제시하고자 한다. 우선 건강상태를 나타낼 수 있는 적절한 지표인 건강사고(health accident)를 정의하고, 개인의 다양한 건강 관련 정보들을 이용해 건강사고 발생가능성을 추정할 수 있는 통계모형을 적합한 후, 이를 통해 유사한 건강리스크에 노출된 사람들의 건강상태를 평가하는 방안을 제시하고자 한다.

본 연구 모형은 개인의 건강상태를 추정하는 통계모형으로 금융회사에서 활용하고 있는 신용평가모형에서 사용하는 로지스틱 회귀모형(logistic regression model)을 이용하였다. 이는 신용정보(상환이력, 신용거래 및 부채 정보 등)를 통해 부도발생 가능성인 신용도를 평점(score) 형태로 측정하는 모형인 신용평가모형이 본 연구에서 제시하고자 하는 개인의 건강관련 정보를 통해 건강사고 발생가능성을 추정하는 기법과 유사하며, 많은 연구를 통해 여러 금융 분야에서 널리 이용되고 있기 때문이다. 또한 Kim 등(2019)에 의하면 로지스틱 회귀모형은 종속변수가 범주형인 신용평가모형에 적용되는 모형으로 모형에 대한 이해가 쉽고, 적합된 결과에 대한 설명이 편리하며, 가변수(假變數, dummy variable)에 대하여 변수의 중요도 파악이나 평점 형태의 분석에 적합하고, 모형의 안정성, 변별력 등에 대한 모니터링에 더 유리하다고 알려져 있다.

본 연구의 결과는 다양한 분야에서 활용될 수 있을 것으로 기대된다. 우선 주요 질병의 진단 및 사망에 영향을 미치는 유의한 변인들과 계량화된 효과를 도출함으로서 건강수준의 향상을 위한 다양한 방안들을 도출할 수 있다. 또한, 건강수준을 평가하고 분류하는 모형은 건강관련 산업 및 서비스 운영에 내재된 리스크를 분석하고 이해하여 새로운 기회를 창출해 내는 데 중요한 정보들을 제공해 줄 수 있을 것으로 기대된다.

논문의 구성은 다음과 같다. 우선 제 2장에서는 건강수준의 평가 및 활용과 이를 위한 모형에 대한 기존 연구들을 소개하였다. 제 3장에서는 건강수준을 평가하기 위한 건강사고의 개념을 정의하고 본 연구의 모형 설계에 활용한 자료 및 자료에 포함된 다양한 변수들을 설명하고 모형에 포함되는 변수들을 범주화하는 과정에 대하여 설명하였다. 제 4장에서는 유의한 설명변수들을 이용하여 건강사고를 추정하는 모형을 적합하는 과정과 결과에 대하여 논의하였고, 제 5장에서는 도출된 모형을 이용하여 건강상태를 점수화하고 해당 점수를 이용하여 건강수준을 분류하는 과정에 대하여 제시하였다. 마지막으로, 제 6장에서는 결론과 함께 연구의 한계점과 후속연구방향에 대하여 간략하게 논의하였다.

2. 선행연구

본 연구는 개인의 건강상태를 나타낼 수 있는 지표를 정의하고, 신용평가모형에서 일반적으로 이용되는 통계기법을 이용하여 해당 지표를 여러 설명변수들을 통해 도출하는 방안을 제시하는데 목적이 있다. 우선 개인의 건강상태를 나타내는 지표로 건강나이에 관한 국내외 선행연구 내용을 참고할 수 있다. Furukawa (1975)는 다중회귀분석을 이용하여 건강나이를 도출하는 방안을 제시하였고, 그 결과를 통해 혈압이 높은 사람의 경우 건강나이가 실제나이보다 높게 나타난다는 것을 확인하였다. Goggins 등(2005)은 중국의 자료를 이용하여 건강나이의 측정을 위해 노화지표를 사용할 수 있음을 제안하였고, Klemera와 Doubal (2006)은 건강나이와 실제나이 간 수리적 관계를 구축하여 건강나이 추정방법을 도출하였다.

Bae 등 (2008)은 여러 신체기능과 관련한 지표들을 바탕으로 건강나이를 추정할 수 있는 방안에 대하여 논의하였고, Park 등 (2009)은 건강나이 도출과 관련하여 임상적으로 유용한 것으로 알려진 바이오마커들을 이용하여 주성분분석을 적용한 건강나이 산출식을 개발하였다. 보다 최근에 이루어진 연구들로, Yoo 등 (2017)은 다양한 바이오마커들의 정보를 근거로 도출된 건강나이가 향후 사망확률 예측에 유용하게 이용될 수 있다는 점을 실증분석하였고, Kang 등 (2018)은 개인의 건강과 노화상태, 사망과 연령에 따른 질병발생 가능성을 예측하기 위해 건강나이를 이용하였다. 또한, Pierleoni 등 (2019)는 신체기능을 측정할 수 있는 장비를 착용하여 장년층에 속하는 사람들의 건강나이를 추정할 수 있는 시스템을 소개하였다.

신용평가모형에서 다양한 요인들을 반영하여 신용도를 나타내는 점수를 도출하는 방법론은 Durand (1941)의 개인 신용도를 평가하여 점수화하는 연구를 시작으로 발전해 왔다. 해당 연구에서는 여러 금융기관의 개인의 신용정보를 이용하여 우, 불량 고객을 구분하고, 판별분석에 따른 주요변수의 가중치를 구하는 방안을 제시하였다. 신용평가모형은 개인이 아닌 기업의 신용도를 평가하는 모형으로도 활용되었는데, Altman (1968), Ohlson (1980), Hamer (1983), Huang 등 (2004)의 연구에서는 판별분석, 로지스틱 모형, 신경망 모형 등 다양한 통계모형을 이용하여 기업의 부실예측모형을 제시하였다.

최근의 신용평가모형 관련 연구로 Hong과 Park (2005)는 의사결정나무 모형을 신용평가모형에 적용하는 방안을 제시하였고, Park과 Kim (2011)의 연구에서는 계층분석과정 방법론을 적용하여 다양한 요인들을 반영한 의료벤처기업의 신용평가모형을 설계하였다. 또한, Jeon과 Seo (2018)는 일반화가속모형을 기술신용평가에 적용하는 방안을 제시하였고, Kim 등 (2019)은 비금융정보인 온라인의 고객 거래정보를 기반으로 로지스틱 모형을 적용하여 개인 신용도 평가를 통한 점수를 도출하는 방안을 제시하였다. 신용평가 모형의 이론적 배경과 모형 설계 과정 및 모형의 활용 등에 관한 내용은 Finlay (2012)에 자세하게 기술되어 있다.

3. 건강수준 지표 및 분석자료

3.1. 건강수준의 측정 지표

개인의 건강수준을 측정하고 평가하기 위해서는 건강수준에 대한 적절한 지표를 결정해야 한다. 어떤 사람의 건강상태는 주요 질병발생 및 사망에 영향을 미치는 중요한 두 요소인 성별과 연령, 그리고 건강과 관련한 행동요소들인 흡연여부, 음주여부, 규칙적 운동여부, 식습관 등의 영향을 받는다. 결과적으로 어떤 개인의 건강상태는 일반적으로 실시하는 건강검진에서 측정하는 다양한 항목(비만도, 혈압, 콜레스테롤 등)들의 결과 수치 및 최근에 이용한 의료기록으로 나타나게 되는데, 그러한 수치들을 종합적으로 고려한 결과를 바탕으로 건강 수준을 측정할 수 있는 지표를 개발해 보고자 한다.

기존의 다양한 분야의 실증연구에서 앞서 언급했던 요소들은 사망률이나 암, 심장질환, 뇌혈관질환과 같은 여러 만성질환에 직, 간접적으로 영향을 미치는 것으로 확인되었다. 본 연구에서는 개인의 건강 수준을 질병에 의한 사망에 더하여 적극적인 치료를 요하는 질병의 발생 가능성을 적절한 통계모형을 이용하여 측정해보고 이를 바탕으로 건강의 수준을 분류할 수 있는 모형을 설계하여 보았다.

구체적으로, 모형화하고자 하는 변수를 설정하기 위해 세계보건기구(WHO)에서 발표한 주요 만성질환인 심혈관질환, 당뇨병, 만성호흡기 질환 및 암의 발병 또는 질병으로 인한 사망을 건강사고로 정의한다. 건강사고의 범위를 보다 명확하게 정의하기 위해 한국표준질병 및 사인분류(Korean Standard Classification of Diseases; KCD)는 통계청 사망원인통계 분류기준을 준용하며, 건강사고는 다음 중 하나 이상의 항목이 발생(진단 또는 사망)하는 경우를 의미한다 (단, 진단의 경우에는 질병의 만성여부를 반영하기 위해 동일한 질병분류 코드가 동일 연도에 2회 이상 나타난 경우만을 대상으로 한다).

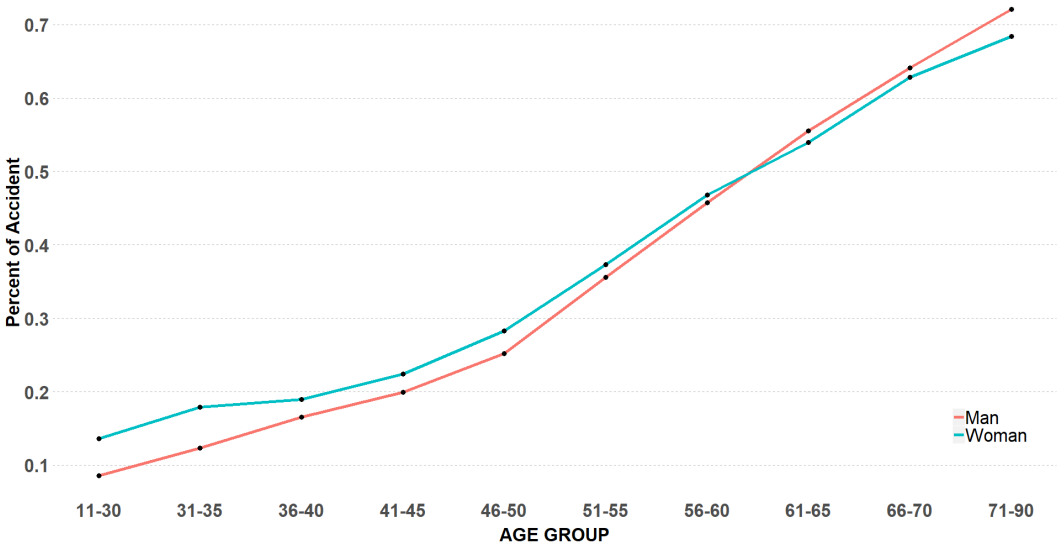

Figure 3.1. Crude rate of health accident by sex and age group.

- 심, 뇌혈관 질환
- 당뇨병
- 만성하부호흡기 질환
- 암 (악성신생물)
- 질병으로 인한 사망 (KCD코드 중 S, T 제외)

실증자료를 이용하여 건강사고에 영향을 미치는 요인들을 탐색해 보고, 해당 요인들에 대한 개인 정보를 바탕으로 건강사고의 발생 확률을 도출하여 개인의 건강 수준을 측정하고 분류하는 모형을 설계하였다.

3.2. 분석자료 및 변수정의

앞 절에서 정의한 건강사고에 영향을 미치는 요인과 건강사고의 발생 확률을 평가할 수 있는 모형(이하 건강 수준 평가모형)의 설계를 위한 자료로의 접근을 위해 S대학교 생명윤리심의위원회(IRB승인번호: SSU-201908-HR-139-01)에서 연구계획 심의면제 승인을 받은 뒤 국민건강보험공단에서 제공하는 표본코호트 DB(연구관리번호: NHIS-2019-2-244) 사용 승인을 받았다. 해당 자료는 국민건강보험의 데이터베이스를 통해 표본 추출된 약 100만명의 건강보험 및 의료급여권자들의 연도별 진료내역, 상병내역, 처방전 내역, 건강검진 내역 등의 의료이용 정보를 포함하고 있는 패널 자료이다. 표본추출은 층별 계통추출법으로 성별, 연령, 소득수준, 지역에 따른 1,476개의 층을 이용하였고, 표본의 대표성을 평가하는 목표변수로 연간 총의료비를 이용하였다. 또한, 표본의 모집단 대표성 평가 결과 대표성이 확보되었다고 평가되었다. 보다 구체적인 표본 설계에 대한 정보는 Lee 등 (2013)을 참조하기 바란다.

본 연구의 분석대상은 2009년과 2010년의 건강검진 수검자 352,896명으로 대상자들의 건강검진 결과에 대한 정보와 자료의 충분성을 확보하기 위해 수검 이후 4년의 기간 동안 건강사고가 발생하였는지의 여부를 관찰하였다. 그 결과 관찰기간 내 총 121,242건의 건강사고(34.4%)가 발생한 것으로 나타났다.

분석대상의 성별, 연령군별 건강사고 발생 비율은 Figure 3.1과 같다. 건강사고의 발생확률은 예상대로 연령에 따라 증가하는 패턴을 보이고 있으며 특히 남성의 경우에는 45세 이후 건강사고의 발생의 증가 속도가 여성보다 빠르게 나타나고 있다. 또한, 60세 이전 연령에서는 여성의 건강사고 발생확률이 높게 나타나고 이후 역전이 일어나는 것으로 관찰되었다.

모형의 설계를 위해 기존 연구결과를 바탕으로 자료에 포함된 건강검진 항목들 중 건강사고 발생에 영향을 미칠 수 있는 요인 (바이오마커) 및 최근 의료이용기록을 고려하였는데, 분석 대상이 된 변수들의 정의와 설명은 Table 3.1에 정리하였다.

모형의 구분과 모형에 포함될 변수를 선택하는 과정에서 신용평가 모형의 변수 선정 및 범주화에서 일반적으로 사용되는 지표인 Weight of Evidence (WoE)와 Information Value (IV)의 개념을 적용하였다. 예를 들어 어떤 변수의 범주화 결과 개의 범주를 얻었다고 하자. i번째 범주 내 포함되는 인원수를 n_i, i번째 범주 내 건강사고가 발생한 인원 수를 e_i라 하고, 전체 인원수를 n_T, 전체 건강사고 발생 인원수를 e_T라 하면 집단 i의 WoE는 다음과 같이 정의된다.

$$(\text{WoE})_i = \ln\left(\frac{n_i - e_i}{n_T - e_T} \Big/ \frac{e_i}{e_T}\right). \tag{3.1}$$

각 범주별로 얻은 WoE값을 이용하여 IV는 다음과 같이 정의된다.

$$\text{IV} = \sum_{i=1}^{k} \left(\frac{n_i - e_i}{n_T - e_T} - \frac{e_i}{e_T}\right) \cdot (\text{WoE})_i. \tag{3.2}$$

일반적으로 IV값이 0.02 이상일 경우, 해당 변수는 모형화하고자 하는 사고의 발생에 대한 오즈비(odds ratio)를 설명하는 데 유용하다고 판단되어 모형의 설명변수로 고려될 수 있음을 의미한다. IV값은 상관도가 높은 여러 변수들 중 하나의 변수를 선택하거나, 범주화 과정에서 IV값을 최대로 하는 범주를 정의하는 데 활용된다.

우선적으로 성인의 경우 성별과 연령에 따라 건강사고의 발생 패턴과 빈도에 차이가 있으므로, 성별과 연령은 모형에 반영되어야 하는 중요한 변수이다. 그러나 각 변수별로 범주화 과정에서 도출되는 IV값을 비교하여 보았을 때, 성별과 연령의 경우 다른 변수들에 비하여 IV값이 상대적으로 매우 높은 값을 갖기 때문에 모형 내 설명변수로 포함된다면 변별력 있는 정교한 모형 개발이 어렵다. 따라서 성별과 연령에 따라서 분리된 집단의 리스크 속성 및 데이터의 충분성 등을 고려하여 성별과 연령을 모형 내 설명변수로 포함하지 않고 모형 구분을 통해 각 성별과 적절한 연령군에 대응하는 모형을 별도로 도출하였다.

각 성별로 동일한 연령군에 포함되는 사람들의 건강 수준을 분류할 수 있는 모형 구분을 위해 건강사고의 패턴이 유사하다고 판단되는 연령군을 재범주화하였다. 구체적으로, 건강검진연도 기준으로 5세 단위로 범주화한 연령군의 자료를 기초로 새로 범주화되는 연령군의 데이터 충분성을 확보하고 (각 성별 전체 인원수의 최소 10%이상) 같은 범주에 포함되는 5세 단위 연령군의 건강사고 발생률이 현저한 차이를 보이지 않는 (10%p 이내) 두 조건을 만족하는 연령군으로 재범주화 한 결과 남성의 경우 4개, 여성의 경우 5개의 연령군을 도출하였다 (Table 3.2).

각 모형별로 설명변수들을 데이터의 충분성 및 건강사고 발생률 등을 고려하여 적절하게 범주화하였으며, 척도가 다른 변수들의 척도 단일화 과정을 통해 해당 변수들의 설명력을 증가시켰다. 이 과정을 통해 이상치(outlier)의 영향을 줄일 수 있었으며, 결측치(missing value)는 하나의 범주로 포함되었다. 또한 설명변수별로 IV값이 작거나, 의학과 역학 분야의 기존연구에서 나타났던 설명력이 뚜렷하게 나타나지 않는 변수들은 고려대상에서 제외하기로 하였으며, 상관성이 높은 변수들의 경우에는 그들 중 IV값이 가장 크게 나타나는 변수를 모형 설계과정에서 고려하기로 하였다.

Table 3.1. Information on the variables obtained from health examination

Inspection method	Measurement	Cohort DB variable name	Analysis variable name	Remarks	Purpose of test
Physical test	Height	G1E_HGHT	HEIGHT	cm	Tested to check for obesity
	Weight	G1E_WGHT	WEIGHT	kg	
	Waist circumference	G1E_WSTC	WAIST	cm	
	Body mass index	G1E_BMI	BMI	Weight(kg) / Height^2(m^2)	
	Systolic BP	G1E_BP_SYS	BP_HIGH	mmHg	Tested to check for high blood pressure
	Diastolic BP	G1E_BP_DIA	BP_LWST		
Urine test	Urine protein	G1E_URN_PROT	OLIG_PROTE_CD	1: Negative(−) 2: Weak Positive(±) 3: Positive(+1) 4: Positive(+2) 5: Positive(+3) 6: Positive(+4)	Tested to check for kidney disease
Blood test	Hemoglobin	G1E_HGB	HMG	g/dL	Tested to check for anemia
	Fasting Blood Sugar	G1E_FBS	BLDS	mg/dL	Tested to check for diabetes
	Total Cholesterol	G1E_TOT_CHOL	TOT_CHOLE		Tested to check for high blood pressure, dyslipidemia and arteriosclerosis
	Triglyceride	G1E_TG	TRIGLYCERIDE		
	HDL-Cholesterol	G1E_HDL	HDL_CHOLE		
	LDL-Cholesterol	G1E_LDL	LDL_CHOLE		
	Serum Creatinine	G1E_CRTN	CREATNINE		Tested to check for chronic kidney disease
	AST (SGOT)	G1E_SGOT	SGOT_AST	U/L	Tested to check for liver disease
	ALT (SGPT)	G1E_SGPT	SGPT_ALT		
	γ-GTP	G1E_GGT	GAMMA_GTP		
Past medical history	Hospitalized Days	VSHSP_DD_CNT, FST_HSPTZ_DT	D_HOSP	Day(s)	Check the medical history of the past year as of the health check-up date
	Medical Care Days	VSHSP_DD_CNT	NUM_CHG	Except Dental(D), Oriental(K), and Psychiatry among Medical and Health Institutions(M)	
	Total Benefit	ED_RC_TOT_AMT	TOT_PAY		
	Prescription Days	TOT_PRSC_DD_CNT	TREAT_D	Day(s)	
	Hospitalized	FST_HSPTZ_DT	HOSP_YN	0: Not Hospitalized 1: Hospitalized	
Question	Smoking	Q_SMK_YN	SMK_STAT_TYPE_RSPS_CD	1: Not Smoking 2: Smoking in the past but quitting now 3: Smoking Still	Ask a question to check for smoking

Table 3.2. Result of regrouping of age groups

Sex	Section	Number of data	Data proportion	Number of health accident	Health accident ratio	Odds	WoE	IV
Male	Under 35	35,627	19.7%	3,790	10.6%	8.40	1.38	0.2671
	35–49yrs	67,227	37.1%	13,865	20.6%	3.85	0.60	0.1168
	50–59yrs	38,891	21.5%	15,392	39.6%	1.53	−0.33	0.0241
	Over 60	39,315	21.7%	25,057	63.7%	0.57	−1.31	0.4141
	Sub total	181,060	100.0%	58,104	32.1%	2.12	-	0.8221
Female	Under 30	18,924	11.0%	2,573	13.6%	6.35	1.30	0.1432
	30–44yrs	44,609	26.0%	9,230	20.7%	3.83	0.80	0.1435
	45–54yrs	46,531	27.1%	15,552	33.4%	1.99	0.15	0.0056
	55–64yrs	32,467	18.9%	16,375	50.4%	0.98	−0.56	0.0624
	Over 65	29,305	17.0%	19,408	66.2%	0.51	−1.22	0.2632
	Sub total	171,836	100.0%	63,138	36.7%	0.65	-	0.5721
Total		352,896	-	121,242	34.4%	-	-	-

자료에서 고려하는 변수들의 상관성을 관찰해 본 결과 체위와 관련하여 비만도를 나타내 줄 수 있는 체질량지수와 허리둘레 간, 혈압의 경우에는 수축기 혈압과 이완기 혈압 간, 콜레스테롤 수치의 경우에는 총 콜레스테롤과 저밀도 지단백(LDL) 콜레스테롤 간, 간기능과 관련한 수치인 AST와 ALT 간 상관성이 높은 것으로 나타났다. 비만도의 경우에는 허리둘레의 결측치가 많이 나타났기 때문에 체질량지수를 고려대상 변수로 선택하였고, 나머지 변수들의 경우에는 앞서 설명한 각 변수별 IV값이 높은 변수를 모형 설계과정에서 고려하기로 하였다.

각 고려대상의 변수들 중 정량적인 수치로 나타나는 연속형 변수들의 경우에는 WoE값이 단조 증가 또는 감소하도록 하는 범주 구분(fine classing)을 먼저 수행하고, 범주별 해당되는 충분한 자료를 확보하기 위하여 처음 구분된 범주들을 재범주화(coarse classing) 하였다. 재범주화 과정에서는 각 범주별로 건강사고 발생건수와 미발생건수가 각각 100건 이상이고, WoE값이 유사하게 나타나는 범주들을 서로 통합하였다. 결과적으로, 다음 장에서 설명하게 될 대표모형에 포함할 대상 변수들을 다음과 같이 결정하였다.

혈압의 경우 남성 대표모형은 이완기 혈압, 여성 대표모형은 수축기 혈압을 고려하였으며, 혈압을 제외하면 공통적으로 체질량지수, 공복혈당, 트리글리세라이드, HDL콜레스테롤, 혈색소, 혈청지피티, 감마지티피, 요단백, 연간 내원일수, 연간 처방일수, 연간 건강보험 급여비용, 흡연여부, 입원여부를 모형에 포함될 후보변수들로 고려하였다. 그리고 앞서 언급한 후보변수들은 단순 로지스틱 회귀분석에서 종속변수인 건강사고 발생과 모두 유의한 관련성을 보였다.

4. 모형의 설계

3장에서 정의된 변수들을 고려하여 각 성별과 연령군에 해당하는 건강사고 발생확률을 추정하고, 유의한 설명변수들을 토대로 건강 수준을 분류할 수 있는 모형을 도출하고자 한다. 이를 위해 3.2 절에서 논의한 바와 같이 남성의 경우 4개의 연령군, 여성의 경우 5개의 연령군에 대한 총 9개의 건강사고 발생확률 모형이 필요하다. 건강사고 발생여부에 대한 자료를 이용하여 건강사고 발생확률 예측을 위한 모형으로 앞에서 언급한 바와 같이 로지스틱 회귀모형을 이용하였다. 모형의 종속변수는 5장에 논의할 평점화를 용이하게 하기 위하여 향후 4년 이내 건강사고가 발생하지 않을 확률을 q로 설정하였으며, 로지스

틱 회귀모형은 다음의 수식으로 표현된다.

$$\ln\left(\frac{q}{1-q}\right) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_p \cdot x_p. \tag{4.1}$$

수식 (4.1) 우변의 $x_1, x_2, \ldots, x_p$는 건강사고 발생확률을 도출하기 위한 설명변수들로 3.2절에서 재범주화되어 고려대상으로 선정된 변수들을 나타낸다. 모든 변수들은 범주화되었으므로 각 변수별로 범주의 수 만큼의 가변수를 정의하되, 종속변수 설정의 경우와 마찬가지로 평점화 과정을 위해 건강사고 미발생 확률을 가장 낮도록 추정하는 범주를 기준 그룹으로 설정하였다. 이후의 논의는 성별, 연령군별로 도출된 총 9개의 모형 중 데이터의 비중이 가장 큰 남성 35–49세 연령군과 여성 45–54세 연령군에 해당하는 대표모형 및 그 결과들을 중심으로 기술하였다.

우선, 모형의 과적합 여부를 판단할 수 있도록 전체 자료 중 70%(남성 47,059명, 여성 32,572명)를 무작위로 추출하여 적합데이터로 사용하고, 나머지 30% (남성 20,168명, 여성 13,959명)의 자료는 모형의 검증을 위하여 자료를 분리(검증데이터)하였다. 이 과정에서 적정한 모형의 성능 평가가 이루어질 수 있도록 분리된 자료 내의 각 성별, 연령군별 건강사고 발생률이 동일하게 유지되도록 분리하였다. 이를 통해 얻은 적합데이터를 이용하여 단계적 변수 선택 방법(stepwise variable selection)을 통하여 유의한 변수들을 도출하였다.

그리고 분류기준값(cut off point)을 결정하는 방법으로 ROC곡선을 이용하였다. 로지스틱 회귀모형으로 도출된 건강사고 발생확률이 정해진 분류기준값 이상일 경우 건강사고가 발생하지 않을 것으로 예측한다고 할 때, 민감도(sensitivity)는 실제로 건강사고가 발생하지 않은 경우에 대해 건강사고가 발생하지 않을 것으로 예측한 상대도수, 특이도(specificity)는 실제로 건강사고가 발생한 경우에 대해 건강사고가 발생할 것으로 예측한 상대도수로 정의한다. 0과 1의 범위에서 결정되는 분류기준값에 따라 민감도와 특이도의 값이 변화하는데, 분류기준값을 매개로 하는 순서쌍 (1 − 특이도, 민감도)의 그래프(receiver operating characteristic (ROC) 곡선)를 나타낸 뒤 가장 이상적인 예측 모형에 해당되는 민감도와 특이도 값이 모두 1인 경우를 나타내는 점 $(0, 1)$에 가장 가까운 ROC 곡선 상의 점에 해당되는 분류기준값을 예측 모형의 건강사고 분류기준값으로 결정하였다 (남성 모형의 경우 0.7937, 여성 모형의 경우 0.6666).

적합된 모형의 성능평가를 위한 지표로 신용평가모형에서 일반적으로 모형의 변별력을 검증하는 지표로 사용되는 area under ROC (AUROC), accuracy ratio (AR) 및 Kolmogorov-Smirnov (K-S) 통계량을 산출하여 확인하였으며, 모형의 성능평가는 적합데이터와 분리된 검증데이터 및 건강사고 발생률을 동일하게 유지되도록 30%를 무작위로 추출한 100개의 샘플데이터를 이용하였다.

AUROC는 ROC곡선의 아래 면적을 의미하며, 그 값이 1에 가까울수록 민감도와 특이도가 높으므로 좋은 분류모형이라 할 수 있다. Hand (2009)에서 AUROC의 값은 일반적으로 0.7 이상일 경우 그 변별력이 좋은 모형이라고 알려져 있다. 그리고 AR은 cumulative accuracy profile (CAP) 곡선을 수치화하여 나타낸 지표로 완전한 모형(perfect model)의 면적과 적합된 모형(rating model)의 면적 사이의 비율을 나타낸다. 신용평가모형의 경우 일반적으로 AR이 0.3 이상이면 변별력이 확보되었다고 판단한다.

또한, K-S 통계량은 두 집단(건강사고 발생자, 건강사고 미발생자)간의 누적분포함수의 차이를 비교하여 동일한 분포에서 나왔는지를 검정하는 지표로서, 두 집단간 분포의 차이, 즉, 누적 건강사고 발생률과 누적 건강사고 미발생률간 차이의 최대 값을 의미한다. 신용평가모형의 경우 일반적으로 K-S 통계량이 0.2 이상이면 변별력이 확보되었다고 판단한다. 대표모형의 적합 결과는 Table 4.1과 Table 4.2에 정리하였으며, 모형의 성능평가 결과는 Table 4.3에 정리하였다.

평가결과 AUROC는 남성 모형의 경우 0.7을 약간 상회하는 수준이며, 여성 모형의 경우 0.7을 약간 하

Table 4.1. Estimated parameters in logistic regression models (male)

Variable name		Description	Estimates	Std. error	Wald statistic	p-value	Proportion of data
Diastolic BP	BP_LWST0	> 102	0.0000	-	-	-	1.4%
	BP_LWST1	≤ 102	0.3036	0.1065	8.12	0.0044	4.9%
	BP_LWST2	≤ 92	0.3839	0.1044	13.51	0.0002	6.0%
	BP_LWST3	≤ 89	0.3959	0.0945	17.55	0.0000	49.0%
	BP_LWST4	≤ 75	0.4170	0.1011	17.01	0.0000	10.1%
	BP_LWST5	≤ 70	0.4334	0.0963	20.25	0.0000	28.6%
	BP_LWST6	Missing	-0.0014	1.6999	0.00	0.9993	0.01%
	Total	-	-	-	-	-	100.0%
BMI	BMI0	> 31.96	0.0000	-	-	-	1.4%
	BMI1	≤ 31.96	0.1586	0.0970	2.67	0.1020	13.5%
	BMI2	≤ 27.43	0.2606	0.1002	6.77	0.0093	9.2%
	BMI3	≤ 26.37	0.2910	0.0948	9.42	0.0022	39.9%
	BMI4	≤ 23.27	0.3551	0.0966	13.51	0.0002	36.0%
	BMI5	Missing	-0.0047	1.2234	0.00	0.9970	0.02%
	Total	-	-	-	-	-	100.0%
Fasting Blood Sugar	BLDS0	> 178	0.0000	-	-	-	1.7%
	BLDS1	≤ 178	0.5424	0.1759	9.50	0.0021	0.5%
	BLDS2	≤ 163	0.9833	0.1426	47.54	0.0000	0.8%
	BLDS3	≤ 148	1.2374	0.1792	47.68	0.0000	0.4%
	BLDS4	≤ 143	1.5130	0.1344	126.79	0.0000	1.0%
	BLDS5	≤ 134	1.6104	0.2571	39.23	0.0000	0.2%
	BLDS6	≤ 133	1.8786	0.1541	148.71	0.0000	0.7%
	BLDS7	≤ 129	1.9988	0.1293	238.92	0.0000	1.2%
	BLDS8	≤ 124	2.1861	0.1253	304.45	0.0000	1.4%
	BLDS9	≤ 120	2.2977	0.1386	274.92	0.0000	1.0%
	BLDS10	≤ 118	2.5729	0.1048	603.20	0.0000	4.7%
	BLDS11	≤ 111	2.7692	0.1003	761.69	0.0000	8.0%
	BLDS12	≤ 105	2.8745	0.1014	803.68	0.0000	7.8%
	BLDS13	≤ 101	2.9207	0.1011	833.81	0.0000	8.2%
	BLDS14	≤ 98	2.9409	0.1000	865.66	0.0000	9.8%
	BLDS15	≤ 95	2.9647	0.0932	1011.26	0.0000	49.3%
	BLDS16	≤ 74	3.0509	0.1153	699.71	0.0000	3.5%
	BLDS17	Missing	3.0172	121.6618	0.00	0.9802	0.01%
	Total	-	-	-	-	-	100.0%
HDL Cholesterol	HDL_CHOLE0	≤ 27	0.0000	-	-	-	0.6%
	HDL_CHOLE1	≤ 37	0.2351	0.1472	2.55	0.1103	9.5%
	HDL_CHOLE2	≤ 44	0.2817	0.1447	3.79	0.0516	20.5%
	HDL_CHOLE3	≤ 55	0.3240	0.1439	5.07	0.0243	35.9%
	HDL_CHOLE4	≤ 65	0.3908	0.1452	7.24	0.0071	20.4%
	HDL_CHOLE5	> 65	0.4260	0.1471	8.39	0.0038	13.2%
	HDL_CHOLE6	Missing	9.9589	97.4456	0.01	0.9186	0.02%
	Total	-	-	-	-	-	100.0%
Liver Function Test	SGPT_ALT0	> 92	0.0000	-	-	-	2.4%
	SGPT_ALT1	≤ 92	0.1061	0.0754	1.98	0.1597	39.0%
	SGPT_ALT2	≤ 28	0.2027	0.0765	7.02	0.0081	40.9%
	SGPT_ALT3	≤ 16	0.2204	0.0809	7.42	0.0064	17.7%
	Total	-	-	-	-	-	100.0%
Urine Protein	OLIG_PROTE_CD0	Positive (+2,+3,+4)	0.0000	-	-	-	0.6%
	OLIG_PROTE_CD1	Positive (+1)	0.2096	0.1511	1.92	0.1654	3.7%
	OLIG_PROTE_CD2	Negative, Weak Positive	0.3439	0.1394	6.08	0.0137	95.5%
	OLIG_PROTE_CD3	Missing	0.1716	0.3153	0.30	0.5862	0.17%
	Total	-	-	-	-	-	100.0%

Continued

Continued

Variable name		Description	Estimates	Std. error	Wald statistic	p-value	Proportion of data
Medical Care Days	NUM_CHG0	> 26	0.0000	-	-	-	4.9%
	NUM_CHG1	≤ 26	0.1716	0.0633	7.34	0.0067	5.3%
	NUM_CHG2	≤ 18	0.2690	0.0670	16.13	0.0001	5.3%
	NUM_CHG3	≤ 14	0.3371	0.0923	13.36	0.0003	1.8%
	NUM_CHG4	≤ 13	0.4467	0.0898	24.76	0.0000	2.0%
	NUM_CHG5	≤ 12	0.5027	0.0721	48.65	0.0000	5.0%
	NUM_CHG6	≤ 10	0.6035	0.0717	70.76	0.0000	6.1%
	NUM_CHG7	≤ 8	0.7017	0.0830	71.50	0.0000	3.6%
	NUM_CHG8	≤ 7	0.8260	0.0814	102.91	0.0000	4.3%
	NUM_CHG9	≤ 6	1.0225	0.0723	200.21	0.0000	10.6%
	NUM_CHG10	≤ 4	1.0613	0.0802	175.13	0.0000	6.3%
	NUM_CHG11	≤ 3	1.3301	0.0806	272.12	0.0000	7.7%
	NUM_CHG12	≤ 2	1.4931	0.0803	345.97	0.0000	9.0%
	NUM_CHG13	≤ 1	1.6050	0.0797	405.43	0.0000	10.5%
	NUM_CHG14	≤ 0	1.8067	0.0757	570.22	0.0000	17.7%
	Total	-	-	-	-	-	100.0%
Total Benefit	TOT_PAY0	$>$ 3,527,660	0.0000	-	-	-	0.6%
	TOT_PAY1	$\leq$ 3,527,660	0.4466	0.1581	7.98	0.0047	0.9%
	TOT_PAY2	$\leq$ 2,088,100	0.5503	0.1252	19.33	0.0000	14.0%
	TOT_PAY3	$\leq$ 322,720	0.5846	0.1299	20.26	0.0000	13.3%
	TOT_PAY4	$\leq$ 171,640	0.5929	0.1318	20.25	0.0000	71.2%
	Total	-	-	-	-	-	100.0%

Table 4.2. Estimated parameters in logistic regression models (female)

Variable name		Description	Estimates	Std. error	Wald statistic	p-value	Proportion of data
Systolic BP	BP_HIGH0	> 138	0.0000	-	-	-	10.3%
	BP_HIGH1	≤ 138	0.1186	0.0468	6.43	0.0112	20.9%
	BP_HIGH2	≤ 126	0.1299	0.0420	9.54	0.0020	64.9%
	BP_HIGH3	≤ 96	0.1459	0.0769	3.61	0.0576	3.8%
	BP_HIGH4	Missing	-0.0596	0.9176	0.00	0.9482	0.02%
	Total	-	-	-	-	-	100.0%
BMI	BMI0	> 27.95	0.0000	-	-	-	8.4%
	BMI1	≤ 27.95	0.0774	0.0489	2.51	0.1132	23.3%
	BMI2	≤ 24.61	0.1898	0.0511	13.81	0.0002	19.6%
	BMI3	≤ 23.12	0.2190	0.0504	18.91	0.0000	24.8%
	BMI4	≤ 21.37	0.3001	0.0519	33.39	0.0000	23.9%
	BMI5	Missing	-0.3049	1.1553	0.07	0.7918	0.01%
	Total	-	-	-	-	-	100.0%
Fasting Blood Sugar	BLDS0	> 121	0.0000	-	-	-	5.1%
	BLDS1	≤ 121	1.2588	0.0823	233.76	0.0000	4.2%
	BLDS2	≤ 111	1.4964	0.0868	297.44	0.0000	3.5%
	BLDS3	≤ 107	1.6980	0.0796	455.31	0.0000	5.5%
	BLDS4	≤ 103	1.7383	0.0671	670.45	0.0000	17.0%
	BLDS5	≤ 96	1.7415	0.0651	716.04	0.0000	25.6%
	BLDS6	≤ 89	1.7644	0.0638	764.51	0.0000	39.2%
	BLDS7	Missing	3.6171	1.8825	3.69	0.0547	0.01%
	Total	-	-	-	-	-	100.0%
Triglyceride	TRIGLYCERIDE0	> 239	0.0000	-	-	-	5.3%
	TRIGLYCERIDE1	≤ 239	0.1178	0.0630	3.50	0.0613	12.3%
	TRIGLYCERIDE2	≤ 157	0.1491	0.0575	6.72	0.0095	41.4%
	TRIGLYCERIDE3	≤ 85	0.2145	0.0595	12.99	0.0003	40.9%
	TRIGLYCERIDE4	Missing	-0.0408	0.3542	0.01	0.9082	0.11%
	Total	-	-	-	-	-	100.0%

Continued

Continued

Variable name		Description	Estimates	Std. Error	Wald statistic	p-value	Proportion of data
HDL Cholesterol	HDL_CHOLE0	≤ 34	0.0000	-	-	-	2.0%
	HDL_CHOLE1	≤ 47	0.1409	0.0890	2.51	0.1134	19.4%
	HDL_CHOLE2	≤ 55	0.1486	0.0889	2.80	0.0944	23.5%
	HDL_CHOLE3	≤ 62	0.1784	0.0898	3.94	0.0471	20.4%
	HDL_CHOLE4	> 62	0.2915	0.0887	10.81	0.0010	34.7%
	HDL_CHOLE5	Missing	-1.1252	1.4360	0.61	0.4333	0.02%
	Total	-	-	-	-	-	100.0%
Liver Function Test	SGPT_ALT0	> 32	0.0000	-	-	-	11.0%
	SGPT_ALT1	≤ 32	0.0762	0.0484	2.48	0.1153	14.7%
	SGPT_ALT2	≤ 23	0.1385	0.0492	7.94	0.0048	14.3%
	SGPT_ALT3	≤ 19	0.1761	0.0483	13.29	0.0003	16.3%
	SGPT_ALT4	≤ 16	0.2302	0.0428	28.97	0.0000	43.7%
	Total	-	-	-	-	-	100.0%
Smoking	SMK_STAT_TYPE_RSPS_CD0	Smoker	0.0000	-	-	-	3.2%
	SMK_STAT_TYPE_RSPS_CD1	Non-Smoker, Past Smoker	0.1620	0.0692	5.47	0.0193	96.0%
	SMK_STAT_TYPE_RSPS_CD2	Missing	0.2977	0.1612	3.41	0.0647	0.8%
	Total	-	-	-	-	-	100.0%
Medical Care Days	NUM_CHG0	> 39	0.0000	-	-	-	7.2%
	NUM_CHG1	≤ 39	0.1109	0.0606	3.35	0.0673	8.1%
	NUM_CHG2	≤ 27	0.1960	0.0666	8.67	0.0032	6.1%
	NUM_CHG3	≤ 22	0.3368	0.0600	31.53	0.0000	13.8%
	NUM_CHG4	≤ 15	0.5086	0.0689	54.49	0.0000	8.3%
	NUM_CHG5	≤ 12	0.6061	0.0666	82.72	0.0000	13.7%
	NUM_CHG6	≤ 8	0.7596	0.0778	95.44	0.0000	7.7%
	NUM_CHG7	≤ 6	0.8148	0.0913	79.65	0.0000	4.4%
	NUM_CHG8	≤ 5	0.9026	0.0946	91.12	0.0000	4.3%
	NUM_CHG9	≤ 4	1.0044	0.1003	100.25	0.0000	4.7%
	NUM_CHG10	≤ 3	1.0636	0.1007	111.46	0.0000	21.7%
	Total	-	-	-	-	-	100.0%
Total Benefit	TOT_PAY0	$> 996{,}520$	0.0000	-	-	-	9.7%
	TOT_PAY1	$\leq 996{,}520$	0.1284	0.0477	7.24	0.0071	19.9%
	TOT_PAY2	$\leq 353{,}460$	0.2935	0.0543	29.24	0.0000	31.0%
	TOT_PAY3	$\leq 125{,}870$	0.4756	0.0705	45.55	0.0000	16.6%
	TOT_PAY4	$\leq 57{,}590$	0.5330	0.0972	30.06	0.0000	8.4%
	TOT_PAY5	$\leq 27{,}760$	0.5655	0.1037	29.73	0.0000	14.4%
	Total	-	-	-	-	-	100.0%

Table 4.3. Testing predictive power of the developed models

		AUROC	AR	K-S
Testing data	Male	0.7159	0.4318	0.3165
	Female	0.6825	0.3650	0.2697
Average of sample data	Male	0.7161	0.4322	0.3175
	Female	0.6876	0.3753	0.2773

회하는 수준으로 유사하게 유의한 수준으로 나타났으며, AR, K-S 통계량은 남자 모형 및 여자 모형 모두 변별력을 확보하는 수준으로서 결론적으로 모형의 적합도가 양호한 수준으로 평가되었다.

추가적으로 건강사고 미발생 확률이 작은 순서대로 자료를 정렬한 뒤 예측된 확률값과 실제 사고 발생여부를 나타내는 그래프를 이용하여 모형의 성능을 평가해 보았다. Figure 4.1의 그래프를 보면 건강사고

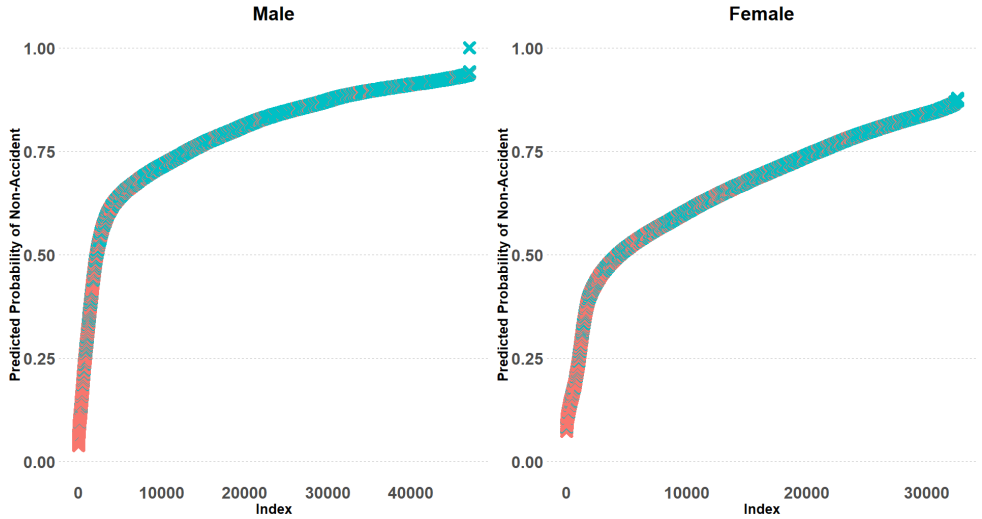

Figure 4.1. Graphical checking of the developed models.

발생확률이 낮게 예측되는 자료일수록 실제로 건강사고가 발생확률이 나지 않은 자료들이 많은 것을 파악할 수 있다. 이는 전반적으로 도출된 모형이 건강사고 발생확률을 적정한 수준으로 예측하는 것으로 볼 수 있다.

5. 건강상태의 평점화 및 분류

4장에서 얻은 모형을 이용하여 개인의 건강관련 정보들을 바탕으로 건강상태를 평점화할 수 있는 방안을 제시하고자 한다. 신용평가모형에서는 다양한 평점화 방법을 고려할 수 있는데, 본 연구에서는 Woo 등 (2013)에서 소개된 평점화 방법 중 오즈를 두 배로 만드는 점수구간의 정의에 따른 points to double the odds (PDO) 방법을 적용하여 건강 수준을 평점화하였다.

전체적으로 건강 수준을 나타내는 점수의 체계는 모형을 통해 도출되는 건강 수준이 가장 높은 경우 (건강사고 발생확률이 가장 낮게 추정되는) 1,000점을 부여하고, 건강 수준이 가장 낮은 경우 (건강사고 발생 확률이 가장 높게 추정되는) 300점을 부여할 수 있도록 설계하였다. 따라서, PDO방법을 이용하여 최고점수는 700점, 최저점수는 0점이고, 50점이 상승하면 건강사고 발생하지 않을 확률에 대한 오즈비율이 두 배가 되는 점수체계를 도출한 후 기본점수 300점을 더하여 평점화하였다.

구체적으로, 특정 성별과 연령군의 모형에서 k개의 유의한 설명변수가 모형에 포함된 경우 각 변수별로 정의된 여러 범주에 해당되는 가변수의 회귀계수 중 가장 큰 값을 $\beta_i^{\max}$라 하자. 4장 로지스틱 회귀모형의 종속변수와 가변수를 정의한 결과에 따라 수식 (4.1)의 우변에서 건강 수준이 가장 좋은 상태는 상수항과 함께 $\beta_1^{\max}, \ldots, \beta_k^{\max}$가 반영되는 경우이고, 건강 수준이 가장 좋지 않은 상태는 상수항만 반영되는 경우이다. 회귀계수는 가변수에 해당하는 범주가 건강사고가 발생하지 않을 확률에 대한 오즈비율에 미치는 영향력을 나타내므로 수식 (4.1)의 우변의 값을 이용하여 평점화하였다. 우선 점수가 50점 상승할수록 건강사고 미발생 확률에 대한 오즈값이 두 배로 상승하게 되는 원점수를 도출하기 위해 각 회귀계수에 $50/\ln 2$의 값을 곱한 뒤 다음과 같이 원점수 S_0를 도출하였다.

$$S_0 = \frac{50}{\ln 2} \cdot \sum_{j=1}^{k} \beta_j^{(\text{obs})}. \tag{5.1}$$

Table 5.1. Classification based on the developed scoring method (male)

Score (S)	Number of data	Proportion of data	Number of health accident	Health accident ratio
$650 \leq S \leq 700$	12,302	18.3%	963	7.8%
$600 \leq S \leq 650$	17,353	25.8%	2,142	12.3%
$550 \leq S \leq 600$	15,048	22.4%	2,692	17.9%
$500 \leq S \leq 550$	12,670	18.8%	3,334	26.3%
$450 \leq S \leq 500$	5,953	8.9%	2,165	36.4%
$400 \leq S \leq 450$	1,500	2.2%	748	49.9%
$350 \leq S \leq 400$	913	1.4%	602	65.9%
$300 \leq S \leq 350$	648	1.0%	471	72.7%
$250 \leq S \leq 300$	411	0.6%	347	84.4%
$200 \leq S \leq 250$	324	0.5%	298	92.0%
$150 \leq S \leq 200$	90	0.1%	88	97.8%
$100 \leq S \leq 150$	15	0.02%	15	100.0%
Total	67,227	100.0%	13,865	20.6%

Table 5.2. Classification based on the developed scoring method (female)

Score (S)	Number of data	Proportion of data	Number of health accident	Health accident ratio
$650 \leq S \leq 700$	5,051	10.9%	772	15.3%
$600 \leq S \leq 650$	7,949	17.1%	1,572	19.8%
$550 \leq S \leq 600$	7,401	15.9%	1,836	24.8%
$500 \leq S \leq 550$	8,084	17.4%	2,523	31.2%
$450 \leq S \leq 500$	7,286	15.7%	2,810	38.6%
$400 \leq S \leq 450$	5,662	12.2%	2,682	47.4%
$350 \leq S \leq 400$	2,483	5.3%	1,354	54.5%
$300 \leq S \leq 350$	736	1.6%	449	61.0%
$250 \leq S \leq 300$	413	0.9%	300	72.6%
$200 \leq S \leq 250$	456	1.0%	364	79.8%
$150 \leq S \leq 200$	551	1.2%	474	86.0%
$100 \leq S \leq 150$	354	0.8%	324	91.5%
$50 \leq S \leq 100$	103	0.2%	90	87.4%
$0 \leq S \leq 50$	2	0.004%	2	100.0%
Total	46,531	100.00%	15,552	33.4%

$\beta_j^{(\mathrm{obs})}$는 어떤 사람의 건강정보를 토대로 j번째 변수에서 관측된 범주에 해당하는 회귀계수이다. 마지막으로 건강 수준이 가장 좋은 상태를 1,000점, 가장 좋지 않은 상태를 300점으로 하는 점수체계 하에서의 점수 S를 다음의 수식을 이용하여 변환하였다.

$$S = (1000 - 300) \cdot \frac{S_0}{\dfrac{50}{\ln 2} \cdot \displaystyle\sum_{j=1}^{k} \beta_j^{\max}} + 300. \tag{5.2}$$

수식 (5.2)에 의하여 변환된 점수를 50점 구간별로 나누어 각 연령군에 포함된 사람들의 건강 수준을 분류하고 각 점수구간별 건강사고 발생 비율의 경험치를 비교해 보았다 (Table 5.1, Table 5.2). 남성과 여성 모두 높은 점수구간에 데이터의 비중이 집중되는 현상을 보이며, 점수가 낮아짐에 따라 건강사고

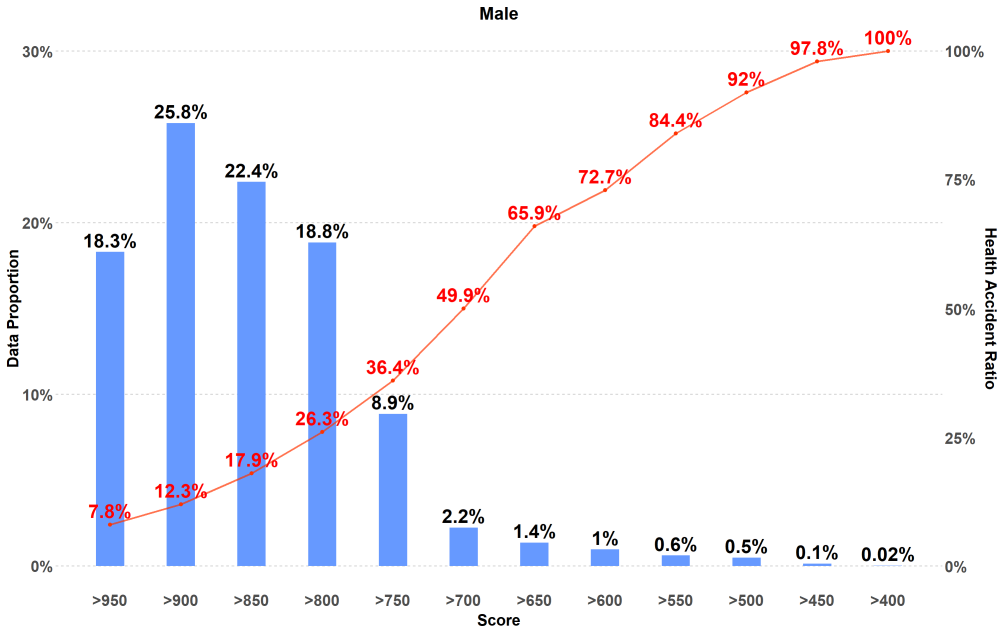

Figure 5.1. Crude rate of health accident calculated by score (male).

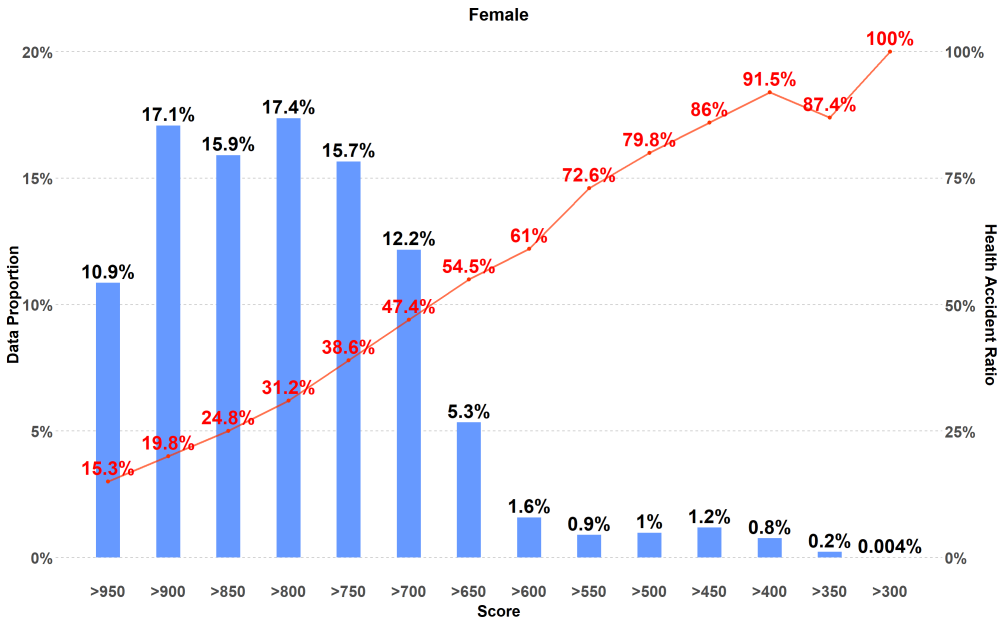

Figure 5.2. Crude rate of health accident calculated by score (female).

발생률이 뚜렷하게 높아지는 것을 확인할 수 있다. 또한, 동일한 점수 구간대의 남성과 여성의 건강사고 발생률을 비교해보면 전반적으로 여성의 건강사고 발생률이 남성의 건강사고 발생률보다 다소 높게 나타나고 있다 (Figure 5.1, Figure 5.2).

이러한 평점화 과정을 통해 각 점수 구간대에 포함된 사람들에게 적합한 건강 증진 프로그램을 설계하여 사람들의 건강관리에 대한 인식을 제고할 수 있을 것으로 기대되며 헬스케어나 의료보험 등에서는 점수 구간대별로 가격을 차등화하거나, 서비스를 이용하는 사람들의 점수 구간별 분포를 활용하여 위험관리에 활용할 수 있을 것이다.

6. 결론

본 연구에서는 건강 수준을 평점화하고 분류하기 위한 모형을 설계하는 방안에 대하여 논의하였다. 건강 수준을 나타내는 지표로 건강사고의 개념을 정의하고 건강사고 발생에 영향을 미치는 요인들을 이용하여 건강사고 발생 가능성을 추정하는 모형을 도출하였으며, 해당 모형을 바탕으로 건강 수준을 평점화하였다. 이를 위해 국민건강보험공단에서 제공하는 표본 코호트 DB를 이용하였고 모형의 설계는 신용평가모형에서 적용하는 방법론을 적용하였다.

표본 코호트 DB에 포함된 건강검진자료 내 여러 고려 대상 변수들을 범주화하여 성별과 연령군으로 구분된 9개의 집단에 대하여 모형을 각각 도출하였다. 건강사고가 발생하지 않을 확률을 로지스틱 회귀모형을 이용하여 추정하고 모형의 예측력을 평가하는 과정에 대하여 논의하였다. 최종적으로 도출된 모형을 이용하여 PDO의 개념을 바탕으로 건강 수준을 평점화하고 분류하는 방안을 제시하였다.

기존 연구결과들을 보면, Katzmarzyk 등 (2012)은 암, 심혈관 및 뇌혈관 질환 등으로 인한 사망위험과 BMI간 유의미한 선형 관계를 확인하였으며, Stocks 등 (2012)은 혈압이 상승함에 따라 암 발생 및 암으로 인한 사망위험이 선형적으로 증가함을 확인하였고, Choi 등 (2018)은 혈압이 상승함에 따라 관련 질병별로 차이는 있으나 전체 사망률과 유의미한 관계가 있음을 확인하였다. 또한, Yi 등 (2017)은 최적 공복혈당(80–94mg/dL) 수치에서 공복혈당이 증가함에 따라 사망위험이 증가하는 관계를 확인하였고, Hernaez 등 (2013)은 ALT 수치가 30U/L 초과해 증가함에 따라 사망위험이 증가함을 확인함으로써 본 연구에서 정의한 건강사고와 고려 대상변수 간 관계와 유사한 결과를 확인할 수 있다.

그러나 Irie 등 (2006)은 크레아티닌이 0.8mg/dL를 초과하여 증가함에 따라 사망위험이 증가함을 확인하였으나 본 연구의 대표모형의 경우 크레아티닌이 증가함에 따라 건강사고의 발생률이 감소하는 경향을 보였으며, Martin 등 (1986)과 Evans 등 (2004)은 LDL 콜레스테롤의 수치가 낮아짐에 따라 당뇨병, 심혈관 및 뇌혈관 질환의 위험의 감소를 확인하였으나, 본 연구의 대표모형의 경우 저밀도 지단백(LDL) 콜레스테롤 수치가 증가함에 따라 위 건강사고의 발생률이 감소하는 경향을 보였다.

본 연구의 결과는 다음과 같은 분야에 활용될 수 있을 것으로 기대된다. 우선 건강에 대한 패러다임은 과거의 질병치료 중심에서 질병의 예방과 건강증진 중심으로 변화하고 있다. 따라서, 본 연구에서 도출한 결과와 같은 개인의 건강 수준에 대한 객관적이고 직관적 해석이 용이한 측정지표를 바탕으로 건강에 영향을 미치는 생활습관을 개선할 수 있는 다양한 프로그램을 설계하고 운영할 수 있다. 이를 통해 사회적으로 건강증진에 대한 의식을 제고시킬 수 있는 결과를 기대할 수 있을 것이다.

또한, 민영의료보험이나 대출 등 건강과 관련된 금융상품 또는 계약에서 건강 수준을 계량화하고 그 결과에 따라 집단을 분류하여 서로 다른 집단에 적용되는 가격 (보험료, 대출이율 등) 을 차등화할 수 있다. 결과적으로, 보다 합리적이고 체계적인 가격 체계를 운영하고 그에 따른 위험관리가 가능해지며 고객에게는 건강개선에 대한 동기를 부여할 수 있을 것이다.

본 연구의 한계점은 다음과 같다. 과거 연구 결과 질병 발병에 영향을 미치는 요인들 중 일부 변수들은 과거 연구에서 나타났던 특성이 나타나지 않는 경우가 발견되었다. 이러한 변수들은 모형에 반영되지 않았는데, 관련 후속 연구들을 위하여 이러한 결과가 나타난 원인을 규명해 볼 필요가 있다. 또한, 본 연구에서는 건강사고 발생 건수를 충분히 확보하여 모형의 신뢰도를 향상시키기 위해 건강사고를 관찰하

기 위한 4년의 추적기간을 설정하였다. 그러나, 건강관련 지표는 시간에 따라 변화하기 때문에 도출된 모형을 이용하면 추적기간 동안 건강관련 지표의 변화를 고려하지 못한다는 한계가 있으며, 데이터에 포함되어 있는 결측치로 인해 적합모형에 대한 결과가 왜곡되어 모형의 적합도 및 요인선택에 영향을 받을 수 있다. 따라서, 보다 충분한 자료를 이용한다면 위 모형을 향상시킬 수 있을 것이다.

그리고 신용평가모형의 경우 판별분석, 신경망모형 및 랜덤 포레스트 모형 등 다양한 모형에 대해 연구가 활발하게 진행되어 각 모형에 대한 추정력 및 안정성에 대한 성능평가가 이루어져 왔으나, 건강 수준에 대한 평가모형은 아직 다양한 통계모형에 대한 성능평가가 이루어지지 않아 이에 대한 추가 연구가 필요해 보인다.

본 연구에서 정의한 건강사고의 개념은 보다 다양한 방식으로 재정의될 수 있다. 보다 합리적인 건강 수준에 대한 지표를 발굴하는 것은 중요한 향후 연구과제가 될 것으로 판단된다. 또한, 향후 축적되는 자료 또는 다른 종류의 자료들을 이용하여 모형을 개선시킬 수 있는 방안에 대한 고민도 필요할 것으로 보인다. 또한, 본 연구에서 도출한 모형을 적용하여 건강 수준이 다른 집단들에 대하여 건강 수준과 관련한 리스크를 세분화하고자 할 경우 적용할 수 있는 구체적인 방안을 제시하는 것도 중요한 후속 연구과제가 될 것이다.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, **23**, 589–609.

Bae, C. Y., Kang, Y. G., Kim, S., *et al.* (2008). Development of models for predicting biological age (BA) with physical, biochemical, and hormonal parameters, *Archives of Gerontology and Geriatrics*, **47**, 253–265.

Choi, J., Jang, J., An, Y., and Park, S. K. (2018). Blood pressure and the risk of death from non-cardiovascular diseases: a population-based cohort study of Korean adults, *Journal of Preventive Medicine and Public Health*, **51**, 298–309.

Durand, D. (1941). *Risk Elements in Consumer Installment Financing* (Technical Ed), National Bureau of Economic Research, New York.

Evans, M., Roberts, A., Davies, S., and Rees, A. (2004). Medical lipid-regulating therapy, *Drugs*, **64**, 1181–1196.

Finlay, S. (2012). *Credit Scoring, Response Modeling, and Insurance Rating: A Practical Guide to Forecasting Consumer Behavior*, Palgrave Macmillan, New York.

Furukawa, T., Inoue, M., Kajiya, F., Inada, H., Takasugi, S., Fukui, S., Takeda, H. and Abe, H. (1975). Assessment of biological age by multiple regression analysis, *Journal of Gerontology*, **30**, 422–434.

Goggins, W. B., Woo, J., Sham, A., and Ho, S. C. (2005). Frailty index as a measure of biological age in a Chinese population, *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, **60**, 1046–1051.

Hamer, M. M. (1983). Failure prediction: sensitivity of classification accuracy to alternative statistical methods and variable sets, *Journal of Accounting and Public Policy*, **2**, 289–307.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Machine learning*, **77**, 103-123.

Hernaez, R., Yeh, H. C., Lazo, M., Chung, H. M., Hamilton, J. P., Koteish, A., Potter, J. J., Brancati, F. L., and Clark, J. M. (2013). Elevated ALT and GGT predict all-cause mortality and hepatocellular carcinoma in Taiwanese male: a case-cohort study, *Hepatology international*, **7**, 1040–1049.

Hong, C. S. and Park, Y. S. (2005). Efficiency comparison of statistical credit evaluation models, *Research Institute of Applied Statistics Sungkyunkwan University*, **13**, 93–107.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study, *Decision Support Systems*, **37**, 543–558.

Irie, F., Iso, H., Sairenchi, T., *et al.* (2006). The relationships of proteinuria, serum creatinine, glomerular filtration rate with cardiovascular disease mortality in Japanese general population, *Kidney International*, **69**, 1264–1271.

Jeon, H. G., Won, J. Y., Peng, X., and Lee, K. C. (2019). Investigating effects of emotional states on the glucose control of diabetes in Korean adults, *Journal of Digital Convergence*, **17**, 301–311.

Jeon, W. J. and Seo, Y. W. (2018). Analysis of important indicators of TCB using GBM, *Journal of Society for e-Business Studies*, **22**, 159–173.

Kang, Y. G., Suh, E., Lee, J. W., Kim, D. W., Cho, K. H., and Bae, C. Y. (2018). Biological age as a health index for mortality and major age-related disease incidence in Koreans: National Health Insurance Service - Health Screening 11-year follow-up study, *Clinical Interventions in Aging*, **13**, 429–436.

Katzmarzyk, P. T., Reeder, B. A., Elliott, S., Joffres, M. R., Pahwa, P., Raine, K. D., Kirkland S. A., and Paradis, G. (2012). Body mass index and risk of cardiovascular disease, cancer and all-cause mortality, *Canadian Journal of Public Health*, **103**, 147–151.

Kim, J. Y., Jang, W. J., and Gim, G. Y. (2019). Development of a personal credit scoring model (COMMERCE Score) using on-line commerce data, *Journal of Information Technology and Architecture*, **16**, 45–55.

Klemera, P. and Doubal, S. (2006). A new approach to the concept and computation of biological age, *Mechanisms of Ageing and Development*, **127**, 240–248.

Lee, J. Y., Kim, K. H., and Lee, J. S. (2013). Construction of Sample Database from National Health Information Database. Seminar on Application of National Health Information Bigdata.

Martin, M. J., Browner, W. S., Hulley, S. B., Kuller, L. H., and Wentworth, D. (1986). Serum cholesterol, blood pressure, and mortality: implications from a cohort of 361,662 men, *The Lancet*, **2**(8513), 933–936.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research*, **18**, 109–131.

Park, C. S. and Kim, M. S. (2011). Credit evaluation model for medical venture business by the analytic hierarchy process, *Asia-Pacific Journal of Business Venturing and Entrepreneurship*, **6**, 133–147.

Park, J., Cho, B., Kwon, H., and Lee, C. (2009). Developing a biological age assessment equation using principal component analysis and clinical biomarkers of aging in Korean men, *Archives of Gerontology and Geriatrics*, **49**, 7–12.

Pierleoni, P., Belli, A., Concetti, R., Palma, L., Pinti, F., Raggiunto, S., Sabbatini, L., Valenti, S., and Monteriù, A. (2019). Biological age estimation using an eHealth system based on wearable sensors, *Journal of Ambient Intelligence and Humanized Computing*, 1–12.

Stocks, T., Van Hemelrijck, M. V., Manjer, J., *et al.* (2012). Blood pressure and risk of cancer incidence and mortality in the Metabolic Syndrome and Cancer Project, *Hypertension*, **59**, 802–810.

Wilson, P. W., Abbott, R. D., and Castelli, W. P. (1988). High density lipoprotein cholesterol and mortality. The Framingham Heart Study, *Arteriosclerosis*, **8**, 737–741.

Woo, H. S., Lee, S. H., and Cho, H. J. (2013). Building credit scoring models with various types of target variables, *Journal of the Korean Data and Information Science Society*, **24**, 85–94.

Yi, S. W., Park, S., Lee, Y. H., Park, H. J., Balkau, B., and Yi, J. J. (2017). Association between fasting glucose and all-cause mortality according to sex and age: a prospective cohort study, *Scientific Reports*, **7**, 1–9.

Yoo, J., Kim, Y., Cho, E. R., and Jee, S. H. (2017). Biological age as a useful index to predict seventeen-year survival and mortality in Koreans, *BMC Geriatrics*, **17**, 7.

건강수준의 측정 및 평점화 모형의 설계

오필재[a] · 김현철[b] · 권혁성[a,1]

[a]숭실대학교 정보통계·보험수리학과, [b]삼정 KPMG Digital 본부

(2020년 1월 7일 접수, 2020년 2월 15일 수정, 2020년 3월 23일 채택)

요 약

최근 기대수명의 증가로 건강에 대한 관심이 늘어나고 있으며 이에 따라 건강관련 산업 및 서비스에 대한 수요도 증가하고 있다. 개인의 건강상태를 다양한 요소들을 이용하여 평가하고 분류할 수 있는 방법을 통해 다양한 건강관련 프로그램 및 서비스를 보다 합리적으로 운영할 수 있을 것이다. 본 연구에서는 기존 연구를 통해 잘 알려진 건강상태 관련 요인들을 이용하여 건강수준을 측정하고 평점화하는 방안을 제시하였다. 이를 위해 신용평가모형의 변수 선정과 범주화, 모형 도출, 평점화로 이어지는 일련의 과정에서 사용하는 방법론을 도입하였고 모형의 적합을 위해서 국민건강보험공단에서 제공하는 표본 코호트 DB를 이용하였다. 본 연구에서 도출된 건강수준 평가모형은 헬스케어 및 건강관련 서비스에 대한 구조 설계 및 운영에 적절하게 활용될 수 있을 것으로 기대된다.

주요용어: 건강수준, 로지스틱 회귀분석, 신용평가모형, 평점화, 표본 코호트DB

[1]교신저자: (06978) 서울특별시 동작구 상도로 369, 숭실대학교 정보통계·보험수리학과.
E-mail: hskwon@ssu.ac.kr