The Korean Journal of Applied Statistics
2022, Vol. 35, No. 4, 457–468

DOI: https://doi.org/10.5351/KJAS.2022.35.4.457
Print ISSN1225-066X / Online ISSN 2383-5818

Development of prediction model identifying high-risk older persons in need of long-term care

Mi Kyung Song[a], Yeongwoo Park[a], Eun-Jeong Han[1,a]

[a]Health Insurance Research Institute, National Health Insurance Service

Abstract

In aged society, it is important to prevent older people from being disability needing long-term care. The purpose of this study is to develop a prediction model to discover high-risk groups who are likely to be beneficiaries of Long-Term Care Insurance. This study is a retrospective study using database of National Health Insurance Service (NHIS) collected in the past of the study subjects. The study subjects are 7,724,101, the population over 65 years of age registered for medical insurance. To develop the prediction model, we used logistic regression, decision tree, random forest, and multi-layer perceptron neural network. Finally, random forest was selected as the prediction model based on the performances of models obtained through internal and external validation. Random forest could predict about 90% of the older people in need of long-term care using DB without any information from the assessment of eligibility for long-term care. The findings might be useful in evidence-based health management for prevention services and can contribute to preemptively discovering those who need preventive services in older people.

Keywords: long-term care, machine learning, prediction model, prevention

1. 연구배경

전세계는 생산가능 인구가 감소하고 노인인구 비중이 증가하는 인구고령화 현상을 경험하고 있다. 우리나라의 경우에도 2018년 고령사회 진입 후 8년만인 2025년에 초고령사회 진입이 예견되고 있으며, 이는 OECD 주요국의 평균 진입 속도 25.08년(OECD, 2021)과 비교할 때, 전세계 유례없는 속도임이 분명하다.

일반적으로 노인은 생체 구조와 기능의 쇠퇴로 인해 질병에 대한 민감성이 높아져 질병 이환에 취약하다. 또한, 질병 및 스트레스 등 외부 영향에 대한 저항력이 감소하여 질병 이환 후 회복이 어려우며, 신체조직과 기관의 퇴행성 변화는 기능저하를 유발하여 결국 기능장애(이하 장애)나 사망을 초래한다 (Nagi, 1991; Verbrugge와 Jette, 1994). 우리나라의 국민이 인지하는 건강수명은 평균 66.3세이며 기대수명은 평균 83.5세로 (Statistics Korea, 2021), 노년기 기간 중 약 17년동안 급/만성 질병 치료를 위한 의료서비스와 기본적인 일상생활수행을 위한 요양서비스를 필요로 하는 기능제한 또는 장애 상태에 놓여있는 것으로 나타났다.

노인의 장애는 요양시설 입소, 공식 · 비공식 재가 서비스 필요와 같은 부정적 결과를 야기하며, 이는 노인 본인뿐만 아니라 비공식 돌봄제공자 및 보건의료 자원에 부담을 주기 때문에 (Thomas 등, 2004), 노인의 장애 발생은 중요한 사회적 문제로 인식되고 있다. 특히, 우리나라 노인장기요양보험 제도 측면에서 노인 수의

This paper was reconstructed from "An exploratory study on risk factors for long-term care needs using the National Health Insurance Service database (Song *et al.*, 2021)"

[1] Corresponding author: Health Insurance Research Institute, National Health Insurance Service, 2 segye-ro, Wonju-si, Gangwon-do, Korea, Republic of. E-mail: 9739han@nhis.or.kr

Published 31 August 2022 / journal homepage: http://kjas.or.kr

증가에 따른 장애 발생률 증가는 장기요양 서비스를 필요로 하는 수급자 증가 및 서비스 수요에 직접적인 영향을 미친다. 장기요양 재정지출 전망에 따르면 장기요양 서비스 이용에 따른 재정지출이 2020년 기준 10.1조원에서 2025년 15.8조원, 2050년에는 116조원으로 급격히 증가할 것으로 예견되고 있다 (Kim, 2020).

최근 WHO는 초고령사회에 보건복지 정책의 실행 목표를 건강노화로 정할 것을 제안하였다. 건강노화란 질병이 있더라도 일상생활수행을 위한 기능적 능력(functional ability)을 최대한 유지하면서 지역사회에서 노년기 웰빙을 가능하게 하는 과정을 일컫는다 (WHO, 2015). 건강노화를 실현하기 위한 방안으로 통합적이고 연속적인 장기요양 체계 구축, 특히, 일차의료 강화를 통한 만성질환 이환 예방, 조기 진단 및 관리를 통한 장기요양 필요(장애발생)의 사전 예방을 강조하였다 (WHO, 2017).

장애의 예방은 크게 질병 예방을 목표로 하는 공중보건 조치(1차 예방), 질병을 치료하는 치유 조치(2차 예방), 손상 또는 장애를 치료하는 치유 및 재활 조치(3차 예방)로 분류된다 (WHO, 1991). 노인의 경우는 이미 대부분 만성질환을 가지고 있으므로, 질병 치료 및 사회적 지지 제공을 통해 삶의 질을 높이고 기능상의 장애를 줄이는 3차 예방이 중요하다 (Won, 2014). 특히, 건강악화를 야기하는 위험요인을 교정하는 예방과 조기 개입은 질병 발생예방뿐만 아니라 더 나아가 건강상태 유지와 건강증진에 중요한 역할을 한다 (Lionis와 Midlov, 2017). 따라서 노인의 생활의존 상태에 영향을 미치는 노쇠, 근감소증, 낙상과 같은 노인증후군의 예방 전략이 필요하며 (Won, 2014), 이러한 서비스 제공 대상인 장애 발생 고위험군을 선별할 수 있는 체계 마련의 중요성이 강조되고 있다.

노인의 비독립성을 야기하는 기능장애 발생 및 장기요양 필요 발생을 예방하기 위한 국가적 노력이 절실하며, 이를 위해 예방 서비스를 개발하거나 제공 체계를 마련하기에 앞서 재원을 효율적으로 활용하고 예방 효과를 극대화할 수 있도록 적절한 대상자를 선별하는 것이 무엇보다 중요하다. 또한, 건강노화 실현을 위한 장기요양 예방 서비스 제공 대상자를 선정하기 위해서는 통상적으로 사회복지서비스에서 활용되는 저소득층, 독거노인 등의 선택적 선별 기준보다는 장기요양 필요를 야기하는 기능장애 발생 고위험 대상자를 선정하는 보편적 기준을 마련해야한다. 건강한 노인이 장애노인으로 이행되는 과정에서 인구사회학적 특성이나 심리적 속성과 같은 개인적 요인, 건강상태, 복지서비스나 사회적, 물리적 환경과 같은 환경(맥락)적 요인 등 노인을 둘러싼 다차원적 영역의 요인들이 서로 상호작용하며 영향을 미친다 (Verbrugge와 Jette, 1994; Freedman, 2009; Yun과 Jo, 2014). 최근 과학적 근거 기반의 실천이나 정책수립의 중요성(evidence based practice & policy; EBP) 및 빅데이터 활용 강화(한국판 뉴딜2.0)가 강조되고 있는 시점에서 국민건강보험공단의 빅데이터에 수집된 전국민의 자격 및 건강정보 등 다양한 요인을 근거로 기능장애 발생 고위험 대상자를 발굴할 수 있는 방안을 마련할 필요가 있다.

그간 장기요양 필요 발생 예방과 관련하여 진행된 연구들은 앞서 언급한 바와 같이 예방의 필요성을 다루는 연구가 주를 이루어, 통계 모형 기반의 예방 정책의 대상자를 발굴하는 것에 초점을 둔 선행연구는 없었다. 다만, 장기요양과 관련하여 장기요양 인정신청을 한 노인의 장기요양 수급 여부를 판단하기 위하여 장기요양등급을 결정짓는 모형을 개발하거나 보완하는 연구(Han 등, 2011; Han 등, 2011)가 있었으나, 이는 신청자를 대상으로 국민건강보험공단 인정조사 직원에 의해 조사되는 장기요양 인정조사표의 정보에 기반하여 구축된 모형이어서 전체 노인의 기능장애 발생 위험을 평가하기 위한 모형으로 적용이 불가능하다.

이에 본 연구는 국민건강보험공단에 집적된 인구사회학적 정보 및 건강정보를 활용하여 타인의 돌봄 및 장기요양 필요를 야기하는 기능장애 발생 가능성이 높은 대상자를 발굴할 수 있는 예측모형을 개발하고자 한다.

2. 연구방법

2.1. 연구설계

본 연구는 국민건강보험공단 국민건강정보를 활용한 후향적 연구(retrospective study)이다. 즉, 2020년도를

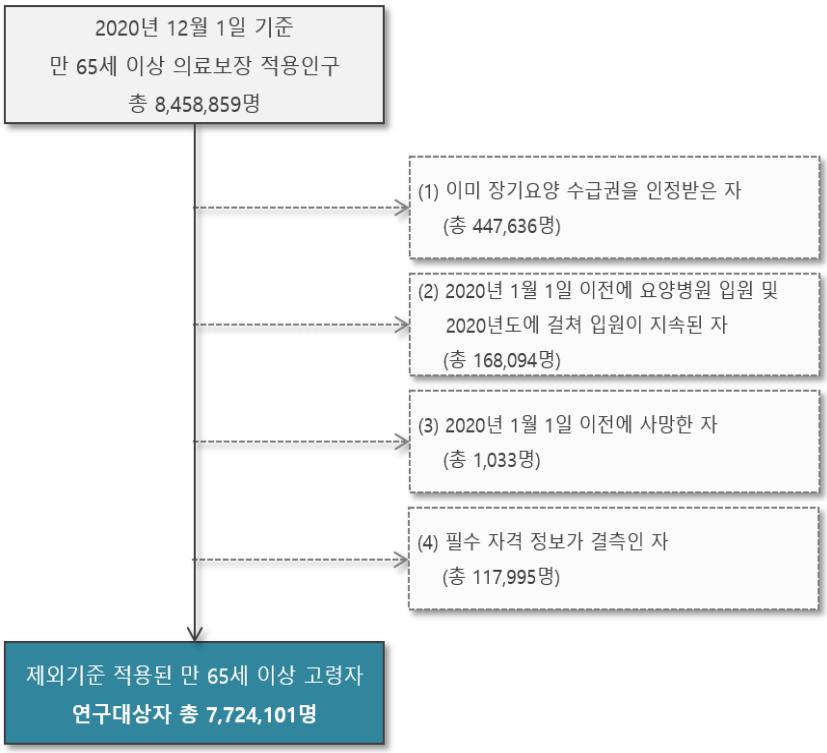

Figure 1: *Study flow chart.*

기준으로 연구대상자를 선정하고, 종속변수의 사건발생 여부를 정의한 후, 사건이 발생한 시점을 기준으로 과거 3년 동안 이미 수집된 국민건강정보를 활용하는 것이다.

선행연구 고찰을 통해 종속변수에 대한 개념적·조작적 정의를 마련하며, 타당한 분석자료 구축을 위해 개념적 틀에 근거하여 선 정의된 종속변수의 사건발생에 영향을 미치는 독립변수를 선정한다. 이를 기반으로 구축된 분석자료를 활용하여 다양한 통계분석 및 체계적 타당성 검증 절차를 통해 예측모형을 개발한다. 본 연구의 윤리적 검토를 위해 국민건강보험공단 건강보험연구원 내 생명윤리심의위원회의 승인을 받았다 (IRB No.: 2021-HR-01-0006).

2.2. 연구대상자

연구대상자는 통상 노인을 정의하는 연령 기준을 준용하여 2020년도 기준 만 65세 이상 고령자로 정의한다. 단, 동등한 상태에서 종속변수를 정의할 수 있도록 이미 타인의 돌봄을 필요로 하는 기능장애가 발생하였다고 판단되는 자는 제외기준으로 정의하였다. 제외기준은 (1) 2020년도 이전에 이미 기능장애로 인한 장기요양 필요가 인정된 노인장기요양보험 수급자인 경우, (2) 2020년도 이전에 요양병원에 입원하여 2020년까지도 입원 상태가 지속되어 돌봄이 필요한 상태로 판단되는 경우, (3) 종속변수를 정의하는 기간인 2020년도 이전에 사망한 경우이며, 이와 더불어 (4) 필수 자격 정보(가입유형, 거주지역, 가구원수)가 결측인 경우도 제외하였다. 조작적 정의 및 제외기준에 따라 국민건강정보 자격 DB를 통해 선정된 연구대상자는 총 7,724,101명이다 (Figure 1).

2.3. 분석변수

2.3.1. 종속변수

기능장애를 측정하는 가장 공통적인 기준은 수단적 일상생활수행능력(Instrumental Activities of Daily Living; IADL)과 일상생활수행능력(Activities of Daily Living; ADL)을 활용하는 것이다(National Council on Disability, 2008). 이는 의학적 진단 이외의 장애 상태에 의해 일상생활 유지 시 장애정도와 타인의 도움 필요정도를 나타내는 사회적 지표이다 (Lim, 2008). 우리나라 노인장기요양보험 역시 이를 고려한 구조화된 등급판정도구를 활용하여 장기요양 서비스 필요 대상자를 정의한다. 등급판정도구는 노인의 주관적인 욕구(want)와 객관적인 필요(needs)를 연계시켜 장기요양 필요의 정도를 판별하는 역할을 한다 (Seo와 Jang, 2005). 즉, 개인의 인지된 주관적 욕구에서 더 나아가 사회제도 내 객관적이고 규범적인 측면에서의 장기요양 필요를 정하고 있다.

전술한 근거에 따라, 본 연구에서 예측하고자 하는 타인의 돌봄이 필요한 의존적 상태인 기능장애 발생은 노인장기요양보험 제도의 장기요양 등급판정절차에 따라 객관적으로 인정된 장기요양 인정여부(장기요양 진입)로 정의하였다. 특히, 노인의 경우 기능상태의 악화가 빠르게 진행되는 취약한 대상자이므로 (Chung, 2015), 단기적으로 가능성이 높은 대상자를 신속하게 발굴할 필요가 있다. 이에 따라 국민건강정보 장기요양보험 DB를 활용하여 2020년도 노인장기요양보험의 수급권을 인정받은 경우(등급판정절차에 의거 장기요양 인정등급을 판정받은 경우)를 종속변수의 사건발생(y = 1)으로 정의하였다. 만일 2020년 12월 31일까지 사건이 발생하지 않은 경우는 사건미발생(y = 0)으로 정의하였다. DB를 통해 조작화된 연구대상자의 사건발생 비율은 4.50%로 관측되었다.

2.3.2. 독립변수

독립변수 선정 및 조작적 정의는 크게 4단계에 걸쳐 마련되었다.

1단계는 독립변수의 선정과정으로 Rowe와 Kahn (1987)의 성공적 노화 개념모델과 Freedman (2009)의 장애노인 모델을 기반으로 Han 등 (2018)이 개발한 노인의 건강노화 개념틀(Figure 2)에 근거하여, 노인의 노쇠 및 기능장애와 관련된 선행연구 고찰(총 42편)과 국민건강정보 상의 가용성 검토를 통해 1차 독립변수를 선정하였다. 그 후 국민건강정보의 자격 DB, 건강검진 DB, 급여내역 DB를 통해 선정된 독립변수의 정보를 수집하여 분석자료를 구축하였다. 이때, 개인적 요소나 환경적 요소 등 시간의 흐름에 변화가 적고 기능장애 발생 직전의 정보가 중요한 독립변수는 최근 1년간의 정보를 활용하였다. 건강행태와 현 건강수준을 측정한 건강검진 자료의 경우 2019년 기준 만 65세 이상 노인의 수검률이 69.9%로 (Statistics Korea, 2021) 약 30%에서 결측이 존재하고, 수검했음에도 항목값에 결측이나 적절치 않은 값이 추가로 발생한다는 점을 고려하여 제외하였다. 이를 대신하여 건강관리를 하고 있음을 나타내는 '건강검진 이력' 변수를 추가하였다. 건강상태를 나타내는 의료이용정보의 경우 Han 등 (2018)에서 제시한 장기요양 노인의 진입 3년 전부터 의료 및 입원이용의 증가 양상 결과를 고려하여 3년 간의 정보를 모두 활용하되, 3년 간의 정보를 분포도에 따른 범주화 (노인성 질환, 의료이용여부, 질환별 수술여부, 수술여부, 질환별 입원여부, 입원여부) 및 평균이나 합계 방식 (다중질환 수, 찰슨동반상병지수, 의료이용일수, 의약품 복용)으로 요약하였다.

2단계는 독립변수의 적절성을 탐색하는 과정으로, 각 독립변수별로 기술통계분석을 통해 결측 정도, 극단치(outlier) 및 오류값 존재여부 등을 검토하여, 각 독립변수의 정의에 맞게 재범주화 등 전처리 과정을 거쳤다.

3단계는 각 독립변수별 가용방안을 모색하는 과정으로 전체 자료의 10%를 단순임의추출한 표본자료를 활용하여, 로지스틱 회귀분석을 통해 얻어진 회귀계수의 유의성(p-값), 예측력(AUC), 정분류율, AIC 통계량을 평가하였다. 이를 통해, 각 독립변수별로 변수의 척도(연속형 vs. 범주형), 범주 수(예: 2범주 vs. 3범주 등)를 결정하였다.

마지막 4단계 과정은 최종 독립변수를 선정하는 것으로 전체 자료에 독립변수별 조작적 정의를 적용한

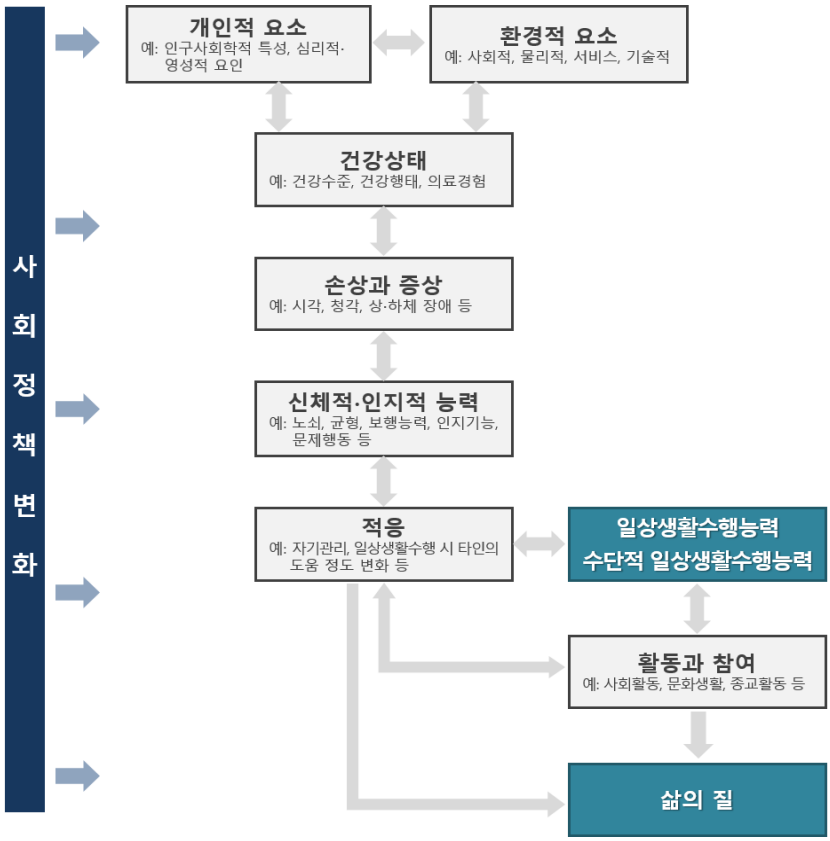

Figure 2: *Conceptual framework.*

후, 로지스틱 회귀분석을 통해 회귀계수의 유의성이 입증되지 않은 변수를 제외하였고, 독립변수 간 다중공선성을 분산팽창인자(Variation Inflation Factor; VIF)로 평가하여 최종 독립변수를 선정하였다.

앞선 일련의 과정을 거쳐 개인적 요소인 인구사회학적 특성 8개, 환경적 요소 중 사회적 환경 1개, 물리적 환경 9개, 건강상태 중 건강수준 9개, 건강행태 1개, 의료경험 49개, 신체적·인지적 능력 1개의 총 78개 독립변수가 선정되었다(Table 1).

2.4. 분석방법

본 연구에서 활용하는 분석자료는 만 65세 이상 전국민 약 770만명에 대한 종합된 건강정보 자료로, 분석자료의 양이 방대하다는 특징을 갖는다. 따라서 예측모형을 개발하기 위한 다양한 통계방법론을 적용할 수 있다. 이에 본 연구에서는 고유한 통계방법인 로지스틱 회귀모형, 머신러닝 방법 중 결과의 시각화로 높은 이해도를 확보할 수 있어 학계 및 산업계에서 많이 활용되는 의사결정나무와 의사결정나무의 과적합 및 불완전성을 보완하기 위해 수많은 결정나무를 만들고 이를 통해 얻어진 결과를 다수결 원칙에 따라 종합하는 알고리즘인 랜덤포레스트, 머신러닝의 최신 기술인 딥러닝 방법의 다층 퍼셉트론 신경망을 예측모형을 개발하기 위한 방법론으로 선정하였다.

각 분석방법별 모형 설정은 다음과 같다. 로지스틱 회귀분석은 고유한 방법 그대로(logit link function)를 이용하며, 유의수준 0.01을 기준으로 하는 후진제거 변수선택법을 이용하였다. 의사결정나무는 이지분리

Table 1: Factors in each domain

개념영역	세부영역	변수	변수 속성	범주 설명
개인적 요소	인구사회학적 특성	성별	이분형	남, 여
		연령	연속형	단위: 세
		직업유무	이분형	무, 유
		소득수준(보험료)	명목형	21범주: 0분위수 ~20분위수
		자격구분	명목형	6범주: 직장가입자, 직장세대원, 지역가입자, 지역피부양자, 의료급여수권자, 의료급여피부양자
		주장애유형[1]	명목형	4범주: 정상, 신체외부장애, 신체내부장애, 정신적장애
		주장애등급	명목형	3범주: 정상, 심하지 않음, 심함
		산정특례대상자 여부	이분형	아니오, 예
환경적 요소	사회적 환경	가족원수	명목형	5범주: 1인, 2인, 3인, 4인, 5인 이상
	물리적 환경	지역구분	명목형	3범주: 대도시, 중소도시, 농어촌
		재정자립도	연속형	단위: %
		사회복지예산	연속형	단위: %
		의료기관 수	연속형	단위: 인구10만명당 기관 개수
		장기요양기관[2]수	연속형	단위: 노인인구 5만명당 기관 개수
건강상태[3]	건강수준	7개 노인성 질환	명목형	4범주[4]: 3년동안 질환없음, 적어도 직전 1년동안 질환없음, 적어도 직전 1년동안 질환있음, 3년동안 질환있음
		다중질환 수	연속형	단위: 최근 3년 연평균 질환 수
		찰슨동반상병지수	연속형	단위: 최근 3년 연평균 CCI
	건강행태	건강검진 이력	이분형	최근 3년내 한번도 없었음, 최근 3년내 한번이라도 있었음
	의료경험	의료이용여부	명목형	3범주: 3년동안 이용안함, 필요에 따라 이용, 3년동안 계속 이용
		의료이용일수	연속형	단위: 최근 3년동안 총 이용일수
		21개 질환별 수술여부 및 수술여부	이분형	3년동안 한번도 수술안함, 3년동안 한번이라도 수술함
		23개 질환별 입원여부 및 입원여부	이분형	3년동안 한번도 입원안함, 3년동안 한번이라도 입원함
		의약품 복용	연속형	단위: 최근 3년 연평균 약물 수
신체적·인지적 능력		노쇠상태	명목형	4범주: 3년동안 노쇠아님, 직전 1년동안 노쇠아님, 직전 1년동안 노쇠상태, 3년동안 노쇠상태

[1] 장애인복지법 시행규칙 제2조의 장애분류 준용함

[2] 요양시설 및 재가서비스를 제공하는 방문요양, 방문간호, 주야간보호, 복지용구 기관에 한함

[3] 건강상태 영역의 변수는 급여DB를 통한 의료이용정보로 정의되며, 이때, 노인장기요양보험법 상 노인성 질환으로 정의된 23개 질환(뇌졸중, 파킨슨, 치매, 우울증, 불면증, 고혈압, 고지혈증, 관상동맥질환, 울혈성 심부전, 당뇨병, 간경변증, 암, 만성기관지염, 천식, 폐결핵, 시각장애, 청각장애, 만성신장질환, 배뇨장애, 관절염, 골다공증, 요통, 골절·탈골 및 사고 후유증)에 대해서만 관측함

[4] 진입 기준 1년 전 구간, 2년 전 구간, 3년 전 구간의 질환별 의료이용 정보를 통해, 구간별 각 질환 유무를 정의하고, 이를 토대로 4범주로 재구성함

를 원칙으로, 엔트로피를 이용하여 독립변수를 선정하고, 향상도를 통해 과적합 여부를 판단하도록 하였다. 또한, 분석을 통해 도출된 노드규칙 정보 및 연구자의 판단, 전문가 의견수렴을 통해 노드 규칙 변경 또는 가지치기를 수행하였다. 랜덤포레스트는 500개의 의사결정나무를 형성하도록 하였으며, 변수 선택에 사용될 변수 중요도는 손실 감소로 설정하였다. 다층 퍼셉트론 신경망의 경우, 분석에 앞서 각 독립변수마다 동일한 학습 영향력을 주도록 0~1 사이의 값으로 표준화하는 작업을 거쳤다. 또한, 직접 연결을 포함한 두 개의 은닉층과

최대 50개의 은닉뉴런수를 갖는 구조로 설정하였으며, 시그모이드 활성화 함수를 이용하였다.

예측모형 개발은 모형 적합 단계와 타당성 검증 단계로 나누어 진행하였다. 우선 전체 자료를 훈련 자료(training data, 60%; n = 4,634,460), 검증 자료(validation data, 20%; n = 1,544,821), 시험 자료(test data, 20%; n = 1,544,820)로 분할하였다. 훈련 자료를 이용하여 각 분석방법별 여러 모형을 적합하며, 검증 자료를 통해 여러 모형 중 가장 성능이 좋은 하나의 모형을 선정하였다. 그 후 시험 자료를 이용하여 각 분석방법별 예측모형의 성능 재현성 검토 및 실질적인 성능을 평가하였다(내적 타당성 평가). 이때, 성능 평가를 위해 모형의 전반적인 예측력을 나타내는 AUC와 모형을 통해 얻어진 예측확률이 얼마나 사건여부를 잘 판별(분류)하는지를 나타내는 판별력(discrimination)을 평가하는 통계량인 정확도(accuracy), 민감도(sensitivity), 양성예측율(positive predictive value), 특이도(specificity)를 이용하였다. 이러한 통계결과는 모형을 통해 얻어진 예측확률을 모형 기반 최적의 절단점(cut-point)을 기준으로 분류한 예측값과 실제 관측된 값 간의 분할표를 기반으로 각 산식에 따라 산출되었으며, 모형별 최적의 절단점은 Youden's index를 통해 결정하였다. 또한, 각 분석방법별 모형이 새로운 자료에서도 적용 가능한지, 일반화 가능성을 평가하기 위하여 외적 타당성 평가를 수행하였다. 이를 위하여 분석자료 구축 절차를 동일하게 적용하되, 연구대상자 및 결과변수를 정의하는 시점을 2021년도 6월 30일 기준으로 업데이트하여 모형구축 자료와는 독립된 새로운 자료를 재구축하였다.

이상의 일련의 결과를 토대로 가장 성능이 적절하다고 판단되는 최종 예측모형을 제시하였다. 모든 통계분석은 SAS 9.4와 SAS E-Miner 14.1을 이용하였다.

3. 연구결과

3.1. 기능장애 발생여부에 따른 독립변수 분포

본 연구대상자의 특성이자 사건발생 여부별 독립변수의 분포를 영역별로 살펴보면, 우선 개인적 요소의 경우, 사건발생 집단에서 여자(71.51%), 낮은 소득수준(특히, 0, 1분위가 24.37% 차지)이 많았으며, 사건미발생 집단에 비해 사건발생 집단의 직업보유률이 5.16%p 낮았다. 사건발생 집단의 평균 연령이 82.45세로 사건미발생 집단에 비해 약 9세 높았으며, 사건발생 집단에서 건강보험 피부양자 자격인 지역세대원(10.96%), 직장피부양자(55.05%), 그리고 저소득층 자격인 의료급여세대주(13.94%)가 상대적으로 많았다. 신체적 장애를 가지고 있는 비율이 32.35%로 높았으며, 산정특례대상자인 경우도 18.87%로 사건미발생 집단에 비해 8.71%p 높았다. 환경적 요소의 경우, 사건발생 집단에서 1인 가구 비율이 36.48%로 높았으며(Δ11.28%p), 농어촌 거주자가 사건미발생 집단에 비해 약 5%p 많았다. 사건발생 집단의 지역별 재정자립도(평균 45.33%)와 사회복지예산(평균 34.02%) 비중이 비교적 작았고, 의료기관수(인구10만명당 평균 106.00개)와 노인인구 5만명당 방문간호기관 수(평균 6.41개)는 상대적으로 적었으나, 요양시설(평균 40.85개), 방문요양(평균 102.52개), 주야간보호(평균 33.10개), 복지용구기관 수(평균 13.24개)는 많았다. 건강상태를 살펴보면, 노인성 질환 중 장기요양 인정자에서 주로 많이 발생한다고 알려진 질환인 뇌졸중, 치매, 고혈압, 당뇨병, 관절염, 골절·탈골 및 사고후유증이 실제로 사건발생 집단에서 높게 나타났다. 특히, 최근 3년동안 고혈압이 있는 경우가 64.00%로 가장 높았고, 관절염(35.40%), 당뇨병(29.32%), 치매(25.11%) 순으로 나타났다. 수발부담이 높은 암의 경우는 오히려 사건미발생 집단에서 이환률이 높으며, 사건발생 집단에서는 직전 1년동안 암에 이환된 비율이 5.45%로 높았다. 최근 3년동안 연평균 다중질환 수는 평균 4.98개로 사건미발생 집단에 비해 약 1개의 질환을 더 많이 보유하고 있으며, CCI 역시 사건발생 집단에서 평균 3.02점으로 약 1점 높았다. 사건발생 집단의 건강검진 수검율은 36.22%로 미발생 집단의 절반수준으로 나타났다. 의료이용여부는 사건발생 집단에서 3년동안 의료를 이용한 비율이 높았으며(94.90%), 의료이용일수 또한 많은 것(평균 총 75.45일)으로 나타났다. 23개 노인성 질환으로 인해 최근 3년동안 한번이라도 수술한 비율은 사건발생 집단에서 34.03%, 사건미발생 집단에서 25.02%였으며, 입원한 비율 역시 사건발생 집단(50.01%)이 사건미발생 집단(27.01%)보다 높게 나타났다. 질환별로는 고혈압, 고지혈증, 시각장애, 청각장애, 배뇨장애를 제외한 모든 수술에서

Table 2: Information of each model and their performance in internal validation

통계방법	예측력	절단점[1]	정확도	민감도	양성예측률	특이도
로지스틱 회귀분석(Logistic regression) 16단계에 걸쳐 총 62개의 위험요인이 선정	0.940	3.0%	85.86	87.96	22.53	85.76
의사결정나무(Decision tree) 총 10개의 위험요인으로 총 19개의 끝마디 생성	0.884	3.6%	89.32	78.18	26.60	89.84
랜덤포레스트(Random Forest) 500개의 결정나무 생성(총 498,179개의 분리규칙)	0.944	5.0%	85.09	90.06	21.87	84.85
신경망·다층퍼셉트론(Multi-Layer perceptron) 78개 독립변수의 직·간접 연결 구조를 갖는 모형	0.948	3.0%	85.22	90.59	22.10	84.97

[1] Youden's Index (J)에 따라 선정된 절단점(cut-point).

Table 3: External validation results of the prediction model

구분	예측력	정확도	민감도	양성예측률	특이도
로지스틱 회귀분석	0.925	84.77	88.36	12.59	84.68
의사결정나무	0.882	88.48	77.85	14.73	88.75
랜덤포레스트	0.942	83.93	90.84	12.25	83.75
신경망 다층퍼셉트론	0.935	84.80	90.38	12.82	84.66

사건발생 집단의 비율이 높았으며, 시각장애, 청각장애, 배뇨장애를 제외한 모든 입원에서 사건발생 집단의 비율이 높게 나타났다. 연평균 복용약물수의 경우 사건발생 집단에서 평균 5.34개로 미발생 집단의 평균 3.51개에 비해 높았다. 신체적·인지적 상태를 나타내는 노쇠율도 사건발생 집단에서 8.08%로 약 4배 높은 것으로 나타났다(별도 표로 제시하지 않음).

3.2. 예측모형 및 내적 타당성 평가

훈련 자료와 검증 자료를 활용하여 분석방법별로 선정된 예측모형의 정보와 이를 시험 자료에 적용하여 산출한 모형별 성능 결과는 Table 2와 같다. 예측력 측면에서 신경망이 94.80%로 가장 좋은 성능을 보였으나, 방법론의 복잡성(complexity)을 감안할 만큼 랜덤포레스트(94.40%) 및 로지스틱 회귀모형(94.00%)에 비해 큰 차이를 나타내지는 않았다. Youden's index 방법을 통해 결정된 각 모형별 최적의 절단점을 이용한 분류표 상에서 얻어진 판별력을 살펴보면, 의사결정나무에서 약 89% 정도의 높은 정확도와 특이도를 보인 반면, 가장 낮은 민감도(78.18%)를 보였다. 의사결정나무를 제외한 모든 모형에서 정확도와 특이도가 약 85% 정도를 보였으며, 신경망이 90.59%로 가장 높은 민감도를 가졌고, 랜덤포레스트도 이에 못지않게 높은 90.06%의 민감도를 나타내었다.

3.3. 외적 타당성 평가

분석방법에 기술한 것과 같이 2021년 6월 1일을 기준으로 외적 타당성 검증 자료를 구축하였다(총 8,080,796명). 이 중 장기요양 진입자는 총 196,868명(2.44%)으로 나타났다. 이 자료를 토대로 각 모형별 외적 타당성 평가의 성능을 Table 3에 제시하였다. 전반적으로 각 모형별 성능의 양상은 유사하나, 모든 모형에서 내적 타당성 평가 결과에 비해 성능이 감소하였다. 상세히 살펴보면, 예측력의 경우 신경망은 93.50%로 여전히 높은 예측력을 보이고 있으나 감소폭이 1.30%p로 비교적 변동이 크게 나타난 반면, 랜덤포레스트는 모형 중 가장 높은 예측력을 보이며(94.20%), 감소폭도 0.20%p로 미미하여 큰 변동없이 성능이 안정적으로 유지되는 것으로 나타났다. 정확도와 특이도는 모든 모형에서 1.20%p 미만으로 감소하였다. 민감도의 경우 의사결정나무와

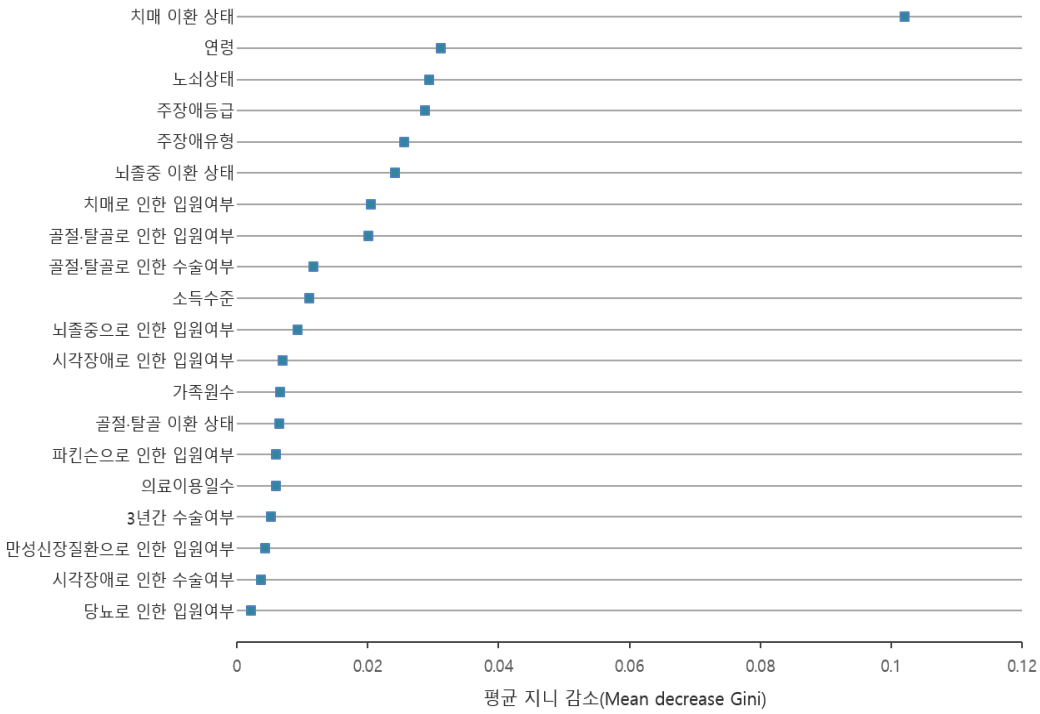

Figure 3: *Top 20 variables by importance.*

신경망에서 성능의 감소를 보였으나, 로지스틱 회귀모형에서 0.40%p, 랜덤포레스트에서 0.78%p 증가하였다.

3.4. 최종 예측모형

전술한 내용을 근거로 모형의 성능이 좋으며, 독립된 새로운 자료에서도 안정적으로 성능이 유지되는 랜덤포레스트를 최종 예측모형으로 선택하였다. Figure 3의 변수 중요도 그림은 평균 지니 감소량이 높은 순서대로 20개의 독립변수를 나타낸 것이다. “3년간 치매 이환 상태”가 장기요양 진입 발생 여부를 분류하는데 가장 중요한 변수로 고려되었으며(총 분리 규칙 개수의 7.9% 차지), “연령(총 분리 규칙 개수의 12.4% 차지)”과 “3년간 노쇠상태(총 분리 규칙 개수의 3.2% 차지)”가 그 다음으로 높은 중요도를 가졌다. 그 외 평균 지니 감소량이 0.02 이상을 가지는 변수는 “주장애등급”, “주장애유형”, “3년간 뇌졸중 이환 상태”, “치매로 인한 입원여부”, “골절·탈골로 인한 입원여부”였으며, 이들 외 나머지 독립변수는 0.02미만의 평균 지니 감소량을 갖는 것으로 나타났다. 최종 예측모형인 랜덤포레스트의 최적 절단점은 5.0%로 나타났다. 추후 활용도를 고려하여, 5~10% 절단점에 따른 예측값과 실제값 간의 분류표 및 판별력 통계 결과를 살펴보았다(Table 4). 절단점이 증가할수록 발생으로 간주되는 대상자가 감소하므로, 정확도와 특이도가 증가하는 반면, 민감도는 감소하는 것으로 확인할 수 있다. 모집단에서의 분포인 4.50%의 2배에 해당하는 절단점 9.00%까지는 민감도가 80% 이상을 유지하였으나, 10.00%부터는 미발생으로 예측되는 대상자의 확대로 높은 정확도(91.65%)와 특이도(91.96%)를 보이는 반면 민감도가 78.91%로 감소하였다.

Table 4: Cross-table and performance by cut-points of random forest model

구분		실제관측값		정확도	민감도	특이도
		장기요양 진입 발생	장기요양 진입 미발생			
절단점 = 5%	발생	178,831	1,280,848	83.93	90.84	83.75
	미발생	18,037	6,603,080			
절단점 = 6%	발생	174,083	1,082,311	86.32	88.43	86.27
	미발생	22,785	6,801,617			
절단점 = 7%	발생	169,333	929,834	88.15	86.01	88.21
	미발생	27,535	6,954,094			
절단점 = 8%	발생	164,603	809,935	89.58	83.61	89.73
	미발생	32,265	7,073,993			
절단점 = 9%	발생	159,903	713,717	90.71	81.22	90.95
	미발생	36,965	7,170,211			
절단점 = 10%	발생	155,352	633,525	91.65	78.91	91.96
	미발생	41,516	7,250,403			

4. 결론

본 연구는 전국민의 인구사회학적 정보 및 의료이용 정보를 기반으로 노인의 기능장애 발생을 예측하는 모형을 개발하여, 장기요양 필요 발생 가능성이 높은 대상자를 선제적으로 발굴하는 것을 목적으로 수행되었다. 본 연구에서는 다양한 통계방법 적용과 체계적인 분석을 통해 타당한 예측모형을 개발하였다. 먼저, 연구자 및 전문가의 판단 하에 단순화된 의사결정나무의 경우, 모형 자체는 적절한 예측력(88.40%), 정확도(89.32%), 특이도(89.84%)를 보였고, 다른 성능에 비해 다소 낮기는 하나 통상 수용 가능한 78.18% 민감도를 보였다. 그러나 타 모형과 비교하여 낮은 성능을 보였다. 복잡한 관계까지도 고려할 수 있는 신경망의 성능이 가장 좋았으나(예측력 = 94.80%, 민감도 = 90.59%), 의사결정나무의 제한점을 보완한 랜덤포레스트와의 성능 차이가 예측력 0.40%p, 민감도 0.53%p로 크지 않았다. 독립변수와 종속변수 간의 일차 선형관계식의 단순한 형태를 갖는 로지스틱 회귀모형도 예측력이 94.00%, 민감도 87.96%로 신경망 및 랜덤포레스트와 유사한 성능을 보였다. 다만, 외적 타당성 평가 결과, 랜덤포레스트가 높은 예측력(94.20%)과 민감도(90.84%)를 보여, 모형 자체의 성능과 유사하게 성능을 유지하는 것으로 나타났다. 이에 본 연구에서는 랜덤포레스트를 최종 예측모형으로 선정하였다. 최종 예측 모형에서 “3년간 치매 이환 상태”의 중요도가 가장 높았으며, 후순위로 “연령”과 “3년간 노쇠상태”가 높게 나타났다. 이는 노인에게 있어 연령과 더불어 치매 이환과 노쇠 발생이 장기요양 필요를 야기하는 기능장애 발생에 가장 영향력 있는 위험요인임을 의미한다. 최종 예측모형을 실제 정책에 활용하기 위한 의미있는 절단점 선정이 필요하며 최적의 절단점인 5%보다 적정 수준의 성능을 유지하는 9%를 활용한다면, 위험요인의 조합에 따른 특정 상태에 놓이는 노인의 경우, 통상적인 상태에서의 장기요양 필요 발생 비율보다 위험도가 2배 높아짐을 시사할 수 있다.

본 연구에서 제안한 예측모형을 국민건강보험공단의 전산시스템에 탑재하여 구현함으로써 예방 서비스가 필요한 대상자를 발굴하는 등 공단이 진행하고 있는 예방 서비스 업무에 활용 가능하다. 이를 통해 발굴된 대상자에게 선제적으로 예방 서비스가 제공된다면 보다 건강한 노후를 지속할 수 있을 것으로 기대할 수 있다. 본 연구는 국민건강보험공단에 집적된 만 65세 이상 전국민 데이터를 활용한 체계적 분석을 통해 예측모형을 개발함으로써 기능장애 발생 가능성이 높은 고위험군을 선별할 수 있다는 데에 의의가 크다. 그러나 노인의 기능장애 발생을 일으키는 수많은 요인 중 행정자료에 국한된 요인만이 활용되어 잠재변인이 있을 수 있다는 한계가 있다. 이러한 한계점을 극복하고 보다 타당도 높은 장기요양 진입 고위험군을 발굴하기 위해 다양한 정보원을 확보하여 개인의 내재적 요인뿐만 아니라 물리적 환경 등의 더 많은 요인을 고려한 후속연구를

수행할 필요가 있다.

References

Chung KH (2015). Characteristic changes and living status of older Koreans, *Health and Welfare Policy Forum*.

Freedman VA (2009). Adopting the ICF language for studying late-life disability: A field of dreams?, *Journals of Gerontoloty: Medical Sciences*, **64**, 1172–1174.

Han EJ, Kwak MJ, and Kan IO (2011). A determining system for the category of need in long-term care insurance system using decision treem model, *The Korean Journal of Applied Statistics*, **24**, 145–159.

Han E, Kwon J, Song MK, Lee Y, Jang H, Kim M, Lee HS, and Kim JH (2018). *Establishing a Prospective Cohort Study of Oder People with Long-Term Care Needs II*, National Health Insurance Service, Wonju.

Han ST, Kang HC, Choi BS, and Lee SK (2011). A study on the judgement rating for level of need for long-term care insurance using a decision tree, *Communications for Statistical Applications and Methods*, **18**, 137–146.

Kim Y (2020). *A Study on the Role Allocation of Private Insurance for Enhancing the Sustainability of Long-Term Care Insurance*, Korea Insurance Research Institute, Seoul.

Lim JG (2008). A Study on the projection of long-term care for the disabled in Korea, *Health and Social Welfare Review*, **25** 185–208.

Lionis C and Midlöv P (2017). Prevention in the elderly: A necessary priority for general practitioners, *European Journal of General Practice*, **23**, 202–207.

Nagi SZ (1991). Disability concept revisited: implications for prevention, *Disability in America. Toward a national agenda for prevention*, 309–327.

National Council on Disability (2008). *National Disability Policy: Progress Report, Chapter 5: Long-term Services and Supports*, National Council on Disability, Washington, DC.

OECD (2021). *OECD health statistics 2021*. https://www.oecd.org/els/health-systems/health-data.htm.

Rowe JW and Kahn RL (1987). Human aging: usual and successful. *Science*, **237**, 143–149.

Seo DM and Jang BW (2005). A study on estimating the cost of long-term care for the elderly by social insurance model in Korea, *Korean Social Security Studies*, **21**, 161–198.

Statistics Korea (2021). *2021 Elderly statistics*.

Statistics Korea (2021). *Life table*. https://www.index.go.kr/unify/idx-info.do?idxCd=4234

Thomas MG, Heather GA, and Theodore RH (2004). Zhenchao G. Hospitalization, Restricted Activity, and the Development of Disability among Older Persons, *JAMA*, **292**, 2115–2124.

Verbrugge LM and Jette AM (1994). The disablement process. *Social Science & Medicine*, **38**, 1–14.

WHO (1991). *Disability prevention and rehabilitation. Executive board. Eighty-ninth session: provisional agenda item 8*. World Health Organization.

WHO (2015). *World Report On Ageing and Health. World Health Organization*. World Health Organization.

WHO (2017). *Global strategy and action plan on ageing and health. Geneva. World Health Organization*. World Health Organization.

Won JW (2014). Health promotion and disease prevention in the older adults, *J Korean Med Assoc*, **57**, 756–762.

Yun E and Jo Y (2014). Factors for environment affecting the social activity disability of the elderly, *Korean Journal of Care Management*, **12**, 21–50.

Received February 24, 2022; Revised April 29, 2022; Accepted May 25, 2022

장기요양 필요 발생의 고위험 대상자 발굴을 위한 예측모형 개발

송미경[a], 박영우[a], 한은정[1,a]

[a]국민건강보험공단 건강보험연구원

요 약

고령인구가 증가함에 따라 국가차원에서 노인의 건강노화 실현을 위한 장기요양 필요 발생의 예방 방안을 마련하는 것은 매우 중요하며, 정책적 효과를 극대화하기 위해서는 적절한 대상자의 선정이 선행되어야 한다. 이에 본 연구는 국민건강보험공단의 국민건강정보를 활용하여, 장기요양 필요를 야기하는 기능장애 발생 가능성이 높은 대상자를 발굴하기 위한 예측모형을 개발하고자 한다. 본 연구는 연구대상자의 과거 수집된 자료를 활용하는 후향적 연구로, 본 연구의 연구대상자는 만 65세 이상 의료보장등록인구이다(총 7,724,101명). 예측모형 개발을 위해 고유 방법인 로지스틱 회귀모형, 머신러닝 방법인 의사결정나무와 랜덤포레스트, 딥러닝 방법인 다층퍼셉트론 신경망을 분석하였다. 체계적 분석절차를 통해 각 분석방법별 모형을 적합하였고, 내적 타당성 및 외적 타당성 평가 결과를 기반으로 최종 예측모형을 랜덤포레스트로 선정하였다. 랜덤포레스트는 모집단에서의 4.50%밖에 되지 않는 장기요양 필요 대상자의 약 90%를 장기요양 필요 발생 고위험 대상자로 예측할 수 있다. 본 연구의 예측모형 및 고위험군 기준은 노인의 욕구 중심에서 예방 서비스가 필요한 대상자를 선제적으로 발굴하는데 기여할 것으로 기대된다.

주요용어: 장기요양, 머신러닝, 예측모형, 예방

본 연구는 송미경 등 (2021)의 "빅데이터를 활용한 장기요양 진입 위험요인 탐색 연구"를 재구성하여 작성하였습니다.

[1]교신저자: (26464) 원주시 세계로 2, 국민건강보험공단 건강보험연구원. E-mail: 9739han@nhis.or.kr